

On Predicting and Compressing Vehicular GPS Traces

Sanjit Kaul and Marco Gruteser
WINLAB, Rutgers University
NJ, U.S.A

sanjit@winlab.rutgers.edu
gruteser@winlab.rutgers.edu

Vinuth Rai and John Kenney
TOYOTA InfoTechnology Center
Mountain View, CA 94043, U.S.A

vinuth@us.toyota-itc.com
johnkenney@alumni.nd.edu

Abstract—Many vehicular safety applications rely on vehicles periodically broadcasting their position information and location trace. In very dense networks, such safety messaging can lead to offered traffic loads that saturate the shared wireless medium. One approach to address this problem is to reduce the frequency of location update messages when the movements of a vehicle can be predicted by nearby vehicles. In this paper, we study how predictable vehicular locations are, given a Global Positioning System trace of a vehicles recent path. We empirically evaluate the performance of linear and higher degree polynomial prediction algorithms using about 2500 vehicle traces collected under urban and highway driving conditions. We find that linear polynomial prediction using the two most recent known locations performs best. Also, traces with a time granularity of 0.2s are highly predictable in low speed urban environments, and a location update rate of 1Hz may suffice to represent urban vehicular movements. Lastly, the paper also evaluates compression of different time-granularity traces using line simplification and polynomial interpolation techniques to reduce message sizes.

I. INTRODUCTION

Vehicle-to-vehicle communications is expected to enable a broad spectrum of safety, traffic management, and infotainment applications [1]. Safety applications, in particular, require that vehicles share their positions and trajectory with other nearby vehicles. This is typically envisioned through periodic broadcasts of Global Positioning System (GPS) coordinates over a Dedicated Short Range Communications channel. It has been shown, however, that the wireless medium represents a bottleneck because it can saturate under dense automotive traffic conditions [2].

While many communications and networking techniques to increase capacity or use available capacity more efficiently have been investigated [3], [4], it is still not fully understood what communication load is necessary to convey vehicular movements in a local region. A typical assumption is that every vehicle needs to transmit positions updates with a frequency of 10Hz to reliably communicate changes in vehicle movement to nearby cars. This update frequency is determined by the rate of change of vehicular positions, the reliability of the communication channel, and application latency requirements.

Recent work has begun to study the required position update frequency in more detail and has demonstrated that vehicular movements can be quite predicable and thus lower transmission rates may be sufficient. If the receiver can predict the path of the transmitter for a certain duration, the transmitter only needs to send the next update when the receivers' prediction becomes inaccurate. In [5] the authors implement such a

scheme through predictive coding techniques with non-linear models to reduce the number of position updates transmitted from each vehicle. The evaluation uses a few sample location traces from highway and urban datasets. In [6] prediction is evaluated using traces from a traffic simulator. To our knowledge, however, there does not exist in the literature a comprehensive study of how well automotive GPS traces can be predicted, for a range of traffic scenarios and driving conditions.

A related question is quantifying the effectiveness of compression techniques for real-world GPS traces. The vehicular communications community is considering to include in vehicle's position update messages not just the most recent position update, but a trace of the n most recent positions. Compression techniques for location traces can therefore help reduce the payload size before transmission.

To address these questions, this paper studies the effectiveness of linear and higher degree polynomial prediction and compression schemes on an extensive set of about 2500 real-world vehicle trips obtained from both a highway and a city environment and with different GPS update frequencies. It concentrates on prediction techniques that do not require a model of vehicle movements. The key contributions are:

- We show that linear prediction based on the two most recent GPS updates outperforms more complex polynomial prediction techniques.
- We show that most prediction errors are caused by incorrect speed estimation, suggesting that using road maps to improve prediction will not yield significant benefits for vehicle traces sampled at high rates.
- We quantify compression gains and show that polynomial compression techniques can outperform the Douglas-Pecker (DP) algorithm [7] for compression of location traces.

The remainder of the paper is as follows. Section II describes the assumptions we make and the different approaches we evaluate. In section III we describe the GPS trace data that we use for evaluations. Section IV presents the results. Related work is described in V. We conclude with a summary of our contributions in section VI.

II. ASSUMPTIONS AND APPROACH

A trajectory/trace is defined as a sequence of one or more locations at points in increasing time. Let (x^t, y^t) represent the location at time t . A trajectory or trace $T(t) = \{x^t, y^t : t = (t_0, t_1, \dots, t_k), t_0 < t_1 < \dots < t_k\}$. We define *trace*

granularity Δ as the time interval between any two adjacent points in $T(t)$ ¹. The trace $T(t)$ stripped of all time information is defined as the available *road/path information* (just the xy co-ordinates, no t).

Typically, at any given time, the location estimate required by a target application needs to be within a certain tolerance of the actual location. The maximum acceptable error is defined as the tolerance δ . At any given time t , it is required that $\sqrt{(\hat{x}^t - x^t)^2 + (\hat{y}^t - y^t)^2} \leq \delta$, where the left hand side of the equation is the Euclidean distance between the location (\hat{x}^t, \hat{y}^t) as estimated by the application, and the actual location (x^t, y^t) .

A. Prediction

Location at a future time is predicted using a certain length of known actual locations, which precede it in time. The actual locations at any time are only available to an *Observer*, for example using a GPS device, and it disseminates (broadcasts) the actual locations in the form of a *location update* to other interested parties. When the prediction error at the interested parties using the last location update exceeds the tolerance, *only then* a new snapshot of actual locations is sent by the *Observer*. In the absence of prediction, location updates will need to be sent at the rate at which a new location is observed. Thus prediction reduces the rate at which the *Observer* needs to broadcast its location.

As the *Observer* needs to know the prediction error at the interested parties, a prediction algorithm will have to run not only at the interested parties but also at the *Observer*.

Prediction using polynomials: Polynomials of various degrees n can be fit to the known actual trace data, exactly or in the least-squares sense (fitting $M + 1 > n$ points). $M + 1$ is the length of known actual locations. The known locations are used to calculate the coefficients, p_0, \dots, p_n , of a polynomial $p(t)$. The coefficients are two-dimensional vectors, (x and y coordinates as elements).

$$p(t) = p_0 + p_1(t - t_0) + p_2(t - t_0)^2 + \dots + p_n(t - t_0)^n \quad (1)$$

The polynomial model thus obtained is used to predict (extrapolate) the location at a time in the future.

We define *linear polynomial prediction* (LPP) as prediction that uses a line model (a polynomial of degree $n = 1$, $M + 1 \geq 2$).

Prediction gains are defined as the reduction in the number of location updates required when using prediction. Specifically, for the evaluation in Section IV, the reduction is defined as the difference between the number of packets sent in absence of any prediction mechanism (one for each point in the GPS trace being evaluated) and the number of packets sent when prediction is used. The gains are shown as percentages

¹The location co-ordinates could be the latitude and longitude or a point's co-ordinates in any other co-ordinate system, for example UTM, which specifies locations on earth as points on a 2-dimensional grid. The times t_0, t_1, \dots don't have to correspond to a fixed sampling rate. However, most GPS devices log time and location information at a fixed rate and hence a fixed Δ .

of the total points in the trace. For a CSMA MAC, such as used in vehicular systems, the reduction in packets will imply equivalent reduction in channel access by the on-road vehicles and hence lesser contention for the wireless medium.

The presented evaluation doesn't consider scenarios where a certain predictor (interested party) may not receive the sent actual locations. For example, a vehicle that enters a road network will not have location information that was broadcast earlier. When prediction is being used, the packets that update the network with the actual locations will contain more than one actual location. The additional locations will make a location update packet larger. However, given the authentication and other header overheads, the increase in packet size can be ignored for the number of actual locations assumed in the evaluation presented. Also, as will be clear in Section IV, a total of two actual locations from the past is most beneficial.

In general, a predictor maybe aided by error corrections or may use prior noisy predictions to come up with a new one. It may also assume models for the process that is being predicted, as in typical Kalman filter based designs. In this work we assume that predictions are made only based on a certain length of actual locations that is known a priori to the predictor. For our study we assume the locations obtained from a GPS device to be the actual locations. We do not account for possible error in location logged by a GPS device.

B. Compression

Compression involves representing a given trace $T(t)$ by fewer points in (x, y, t) space, as long as the error in location estimated from the compressed trace, at any given time, is within the tolerance δ of the actual. Under prediction only the current actual location and the past is known. Compression on the other hand has the entire trace information a priori.

Compression achieved is defined as the difference in the points in the actual trace and its compressed version. Gains are shown as percentages of the total number of points in the trace evaluated.

1) *DP*: The Douglas-Peucker (DP) Algorithm approximates a trajectory by fewer lines. Given any two points in the trajectory, the algorithm tries to minimize the number of lines that connect the points, while ensuring that the error tolerance requirements are satisfied [7].

2) *Polynomial Interpolation (PI)*: We use polynomial approximation of a car's trajectories. Let the functions $x(t)$ and $y(t)$ represent respectively the x and y co-ordinates, of a trajectory, changing with time.

In Equation 1, $p_k, k = 1, \dots, n$ are the polynomial coefficients, and p_0 is the value of the co-ordinate at time t_0 . The number of co-ordinate points encoded by a polynomial approximation is $M + 1$, where $M \geq n$. Let $p_x(t)$ and $p_y(t)$ be the polynomials approximating $x(t)$ and $y(t)$ respectively, for $t \geq t_0$. At any time t , $\sqrt{(p_x(t) - x(t))^2 + (p_y(t) - y(t))^2} \leq \delta$ needs to be satisfied.

The above condition is satisfied by ensuring that $|p_x(t) - x(t)| \leq \frac{\delta}{\sqrt{2}}$ and $|p_y(t) - y(t)| \leq \frac{\delta}{\sqrt{2}}$ are satisfied at all t .

3) *Distance-Time*: The Distance-Time (DT) algorithm first encodes the xy-part of the trace $T(t)$ using minimum number of line segments. It then encodes the distance travelled along each of the line segments, using a line simplification approach (as in DP), as constant velocity segments.

III. TRAJECTORY DATA

We evaluate aforementioned prediction and compression schemes using GPS traces collected by on-road vehicles. Specifically, we answer the following questions:

- What prediction gains are achieved for polynomial models of different degrees n and lengths of known actual locations $M + 1$?
- How are the gains affected by the tolerance δ ?
- How are prediction gains affected by trace granularity Δ ?
- Is a certain Δ smaller than needed in a given environment, for example city or highway roads?
- By how much does knowing road information improve prediction?
- At what speeds do we get prediction gains and what distances can be predicted on average?
- How well do different schemes compress traces as a function of Δ ?

The traces can be organized into three sets on the basis of the frequency at which the GPS information was logged.

A set of 34 traces recorded in and around a downtown in New Jersey were logged at 5Hz^2 , i.e., $\Delta = 0.2\text{s}$. The average speeds in the collected traces were less than $40\text{mph} \approx 60\text{km/h}$. The total trace length is about 30 hours.

Another set of 134 traces, collected in New Jersey, by four drivers over their daily commutes, were logged at 1Hz , i.e., $\Delta = 1\text{s}^3$.

Finally, two sets of traces, one from 20 cars driving over a 4mile highway section in Oakland CA and the other from 100 cars going back and forth over a highway section were collected at a time granularity of 3s . The two sets together are a total of about 2300 traces and more than 400 hours in time.

The GPS latitude and longitude are converted into Universal Transverse Mercator co-ordinates, UTM-x and UTM-y, for all evaluation that follows.

IV. RESULTS

A. Prediction

1) *Prediction gains for different polynomial degrees and number of known actual locations*: Figure 1(a) shows the prediction gains obtained for $\delta = 0.5\text{m}$ and varying degree n of the extrapolating polynomial⁴. In the plot we choose $M = n$, where $M + 1$ is the total number of known actual locations. For a fixed n , choosing $M > n$ leads to lower prediction gains than $M = n$, however. Traces with $\Delta = 0.2\text{s}$ were chosen for the plot. The gains are the greatest for

$\Delta = 0.2\text{s}$ (median of $\approx 80\%$ for $M = n = 1$) and fall considerably for granularities of 1s and 3s (not in plot) to a median of $\approx 30\%$ and $\leq 10\%$ respectively. Also, for any Δ , the combination of $M = n = 1$, i.e., linear extrapolation using two actual locations from the past, provides the maximum prediction gains. The gains, shown for $n \leq 4$, reduce as the degree n increases. Last but not the least, the lack of gains at larger trace granularities, suggests a minimum rate at which location may need to be broadcast. Since these results show that LPP outperforms higher order polynomial prediction, the following results focus on LPP.

2) *Gains using LPP*: LPP achieves median gains of 80% , see Figure 1(b), for $\delta = 0.5\text{m}$ and traces with a time granularity of $\Delta = 0.2\text{s}$, which were taken in a city environment (refer Section III). Gains of 80% suggest that, with prediction, only a packet per second may be sent to update location, instead of a packet every 0.2s .

The figure also shows the improvement in prediction gains, for LPP, obtained with increasing δ . Median gains for a tolerance of $\delta = 1\text{m}$ are 6% , 10% and 15% higher than for $\delta = 0.5\text{m}$, for Δ of 0.2 , 1 and 3s respectively. For tolerances greater than 1m even higher prediction gains were observed. The improvements are most notable for the larger Δ . For $\Delta = 1\text{s}$ the gains at $\delta = 2.0$ and $\delta = 4.0$ were 67% and 77% respectively. For $\Delta = 3\text{s}$ they were 31% and 48% respectively.

3) *Benefits of road information*: LPP implicitly assumes that the direction of motion and speed in the future are the same as the current known direction and speed, which are calculated from the known actual locations. Errors in prediction will therefore occur on a change of either or both and are more likely to occur for larger Δ . Errors due to changing direction alone, can be eliminated if the predictor has access to *road information* (RI), which is the xy part of the actual future trace.

We will show that road information is of very limited benefit in prediction.

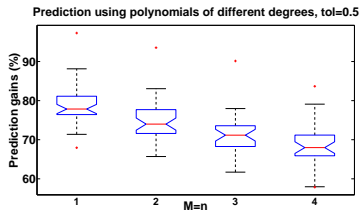
We define xy-compressibility of a trace as the reduction achieved in the number of straight lines that can approximate a given trace, considering only the xy-plane, i.e., we ignore the time co-ordinate of the trace. The greater the percentage compressibility, the lesser the number of straight lines required in proportion to the total in the uncompressed trace. Compressibility of a trace in xy provides a rough estimate of how well direction can be predicted, using LPP, along the trace. In the absence of changing speed, LPP over a trace should achieve prediction gains in the range of the trace's compressibility. This is because, at constant speed the errors in prediction will occur only when the road changes direction, i.e., at all those points in the trace where its compressed version adds a new line (or direction).

Figure 1(c) shows the xy-compressibility of the traces, grouped by their granularity Δ . A tolerance of $\delta = 0.5\text{m}$ is assumed. The granularities of 0.2 , 1 and 3s have median compressibility of 94% , 64% and 52% respectively. The achieved prediction gains for $\delta = 0.5\text{m}$, Figure 1(b), are 80% , 40% and 8% respectively. LPP unable to achieve prediction

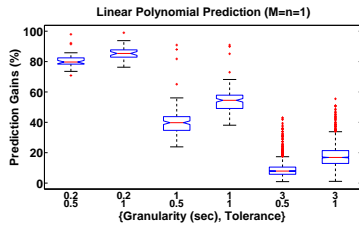
²Garmin GPS 18 device was used

³The Holux GPSlim 236 GPS device was used

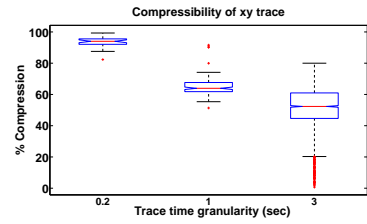
⁴The plot shows results for extrapolation done independently along (x, t) and (y, t) . The basic trends are similar for joint (x, y, t) extrapolation.



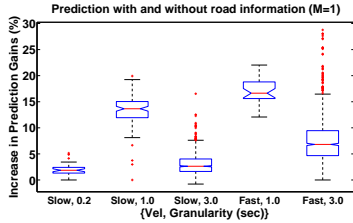
(a) Prediction Gains for polynomials of various degrees.



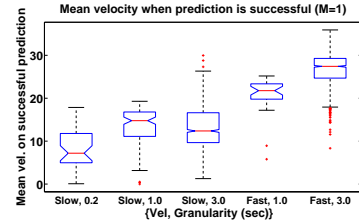
(b) Linear Polynomial Prediction Gains: x-axis first row is granularity Δ s, 2nd row is tolerance δ m



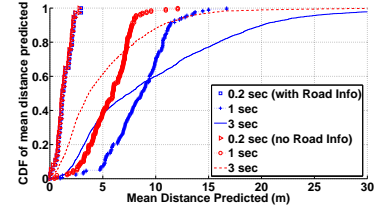
(c) Compressibility of xy-traces.



(d) Improvements in prediction gains on using road information.



(e) Successful prediction occurs at slower velocities (shown in mph).



(f) Distribution of mean distance predicted with and without road information.

Fig. 1. Prediction of location using GPS traces

gains close to the xy-compressibility can be explained by its inability to predict changes in speed.

Consider again $\Delta = 1$ s and $\delta = 0.5$ m. The xy-compressibility is 64%, i.e., 36% of the trace has unpredictable changes in direction. If speed was constant, the presence of RI would lead to prediction gains of 100%, as RI would lead to a priori knowledge of all changes in direction.

Figure 1(d) shows the improvement in prediction gains using LPP, when the predictor has RI. The increase in prediction gains is the difference between prediction gains when RI is known and when it isn't. It is plotted for all granularities. Further, for each time granularity the traces are grouped by their mean speed⁵. Speeds > 41 mph ≈ 65 km/h is labelled as *fast*. The speeds ≤ 41 mph are labelled as *slow*⁶.

The improvement in prediction gains is a median of $\approx 2.5\%$ for $\Delta = 0.2$ s traces, smaller than that at Δ of 1 and 3s. The 0.2s trace benefits very little from RI as it is already very close to xy-compressibility (94%) and could have benefited a maximum of 6%. The gains being less by $\approx 3.5\%$ may be attributed to changes in speed entangled with that of direction.

For the case of $\Delta = 3$ s, prediction seems to be greatly impaired by lack of speed prediction. Although an improvement of 6.8% is seen at faster speeds, xy-compressibility of 52% suggests that an improvement of around 48% is possible in the absence of speed errors. Lastly, faster speeds benefit more from RI (as seen for $\Delta = 1$ and 3s in Figure 1(d)). This can be explained by the more frequent changes in direction at faster speeds, which given the RI don't contribute to errors in prediction.

To summarize, RI is of very little benefit at $\Delta = 0.2$ s. The

⁵Speed between two consecutive points in a trace is calculated as the distance between the points in the xy-plane divided by their separation in time.

⁶The selection of 41mph gives equal number of traces with fast and slow mean speeds.

benefits at Δ of 1 and 3s are greater. However, the prediction gains obtained are still limited due to speed prediction errors.

4) *Speeds and distances at which prediction gains were obtained:* Figure 1(e) shows the distribution of mean speeds over time intervals of successful prediction. Even for the traces with *fast* mean speed, the mean speed during successful prediction is in the *slow* range (medians are < 30 mph). Successful prediction at Δ of 1 and 3s is limited to slow speeds.

Lastly, in Figure 1(f) we plot the cumulative distribution of the mean predicted distance⁷ for the traces of different granularities. For $\Delta = 0.2$ s, 80% of the times (80th percentile) the mean is ≤ 2 m. Road information changes the mean predicted distance by a negligible amount. The 80th percentile improves by 8m for traces with $\Delta = 3$ s and by 3.5m for the traces with $\Delta = 1$ s. The median for $\Delta = 3$ s is less than that for traces with $\Delta = 1$ s, further evidence that prediction gains for larger time granularity traces are impaired by speed prediction errors and the gains achieved are mostly at slow speeds.

The increase in the payload of a location update packet is negligible on using LPP with $M = n = 1$. A network of cars that uses LPP will have to run as many predictors as the number of cars in the network. However, LPP involves extrapolating along a straight line and hence is computationally simple.

B. Compression

We evaluate the percentage compression achieved using polynomial interpolation (PI), DP and DT algorithms (Section II). The compression gains for coarse grained topologies are shown in Figure 2. The gains are evaluated over the traces collected at $\Delta = 3$ s (Section III). The allowed error tolerance

⁷The calculation of the mean assumes the predicted distance to be 0 when prediction along a trace fails.

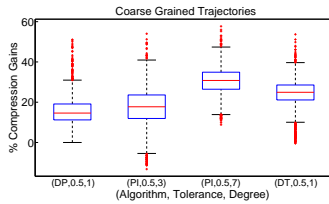


Fig. 2. Percentage compression for coarse grained topologies.

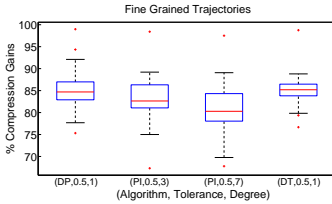


Fig. 3. Percentage compression for fine grained topologies.

is fixed at $\delta = 0.5\text{m}$. In the plot we show gains on using DP, DT and PI using degrees of 3 and 7. PI using degree 7, achieves maximum gains (a median of $\approx 30\%$). The next best gains are achieved by DT, a median of $\approx 25\%$. DP achieves a low of 15% gains. For coarse grained traces, evaluated over different tolerances, polynomials of degree 6-7 gave maximum compression gains.

Figure 3 shows the compression gains achieved for the traces with $\Delta = 0.2\text{s}$. DP and DT do equally well at compressing the traces, achieving median gains of $\approx 85\%$. PI, especially using degree $n > 3$, achieves a tad lesser compression. Given the much greater time complexity of PI, it may not be preferred for compressing fine-grained traces. DT on the other hand does well compressing both fine and coarse grained traces. However, its time complexity is greater than DP.

V. RELATED WORK

Predictive coding is shown to reduce channel load in [5]. The authors a priori assume a Newtonian model for a vehicle. A Kalman filter update scheme with noiseless observations is used. Predictive coding gains are evaluated using two urban and one highway datasets. In [6] the authors use a Kalman filter estimator approach supported by a variable rate communication scheme to reduce the rate at which safety messages need to be sent.

In our work we evaluate the predictability of location based on the availability of a certain length of trace of GPS coordinates from the past. The model used for prediction is extracted from the GPS trace.

Studies, such as [3], [4], [8], [9], suggest mechanisms at the medium access control (MAC) layer to support high offered load in vehicular networks, instead of coding at each node to reduce the offered load.

Compression of spatio-temporal trajectories is studied in [7]. The compressed trajectories should be able to support different database query types. The Douglas-Pecker algorithm is compared with that of Haar wavelets method.

In our work we compare the compression gains from Douglas-Pecker algorithm with polynomial interpolation for different trace granularities. An algorithm, that compresses trajectories in xy-plane and then uses line simplification to encode changes in velocity is also compared to DP and polynomial interpolation.

VI. CONCLUSIONS

We compared polynomial based schemes for the prediction and compression of GPS trajectories. The schemes were evaluated over a large GPS trace data set of ≈ 2500 traces, collected in urban and highway environments. Different GPS trace time granularities were evaluated. Specifically, we conclude

- Linear polynomial prediction using two most recently known locations, gives the maximum prediction gains.
- A location update rate of 1Hz may suffice for vehicles on city roads. GPS traces collected with mean velocities of $< 40\text{mph}$, and trace granularity of 0.2s show a high predictability of 80%.
- Road information has very limited benefit for prediction in city conditions when predicting a 0.2s trace granularity. At granularities of 1s and 3s road information is beneficial, however the additional gains achieved are much lower than expected from xy-compressibility.
- At time granularities of 1s and 3s prediction is only successful at low speeds. Traces collected in highway conditions, with mean speeds $> 40\text{mph}$, showed prediction gains only during stretches when mean speeds were $< 30\text{mph}$.
- We compare compression obtained using polynomial interpolation (PI), a distance-time (DT) approach, and Douglas-Pecker (DP). DT performs well across granularities. It does better than PI for the traces at 0.2s granularity. It performs better than DP by a median of 10% for traces with granularities of 1 and 3s.

REFERENCES

- [1] "Vehicle safety communications project task 3 final report." [Online]. Available: <http://www.nhtsa.dot.gov/staticfiles/DOT/NHTSA/NRD/Multimedia/PDFs/Crash%20Avoidance/2005/CAMP3scr.pdf>
- [2] J. Yin, T. ElBatt *et al.*, "Performance evaluation of safety applications over DSRC vehicular ad hoc networks," in *Proceedings of the 1st ACM international workshop on Vehicular ad hoc networks*. ACM, 2004.
- [3] T. K. Mak, K. P. Laberteaux *et al.*, "Multichannel medium access control for dedicated Short-Range communications," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 1, pp. 349–366, 2009.
- [4] R. Mangharam, R. Rajkumar *et al.*, "Bounded-Latency alerts in vehicular networks," in *2007 Mobile Networking for Vehicular Environments, 2007*, pp. 55–60.
- [5] C. Robinson, D. Caveney *et al.*, "Efficient message composition and coding for cooperative vehicular safety applications," *Vehicular Technology, IEEE Transactions on*, vol. 56, no. 6, pp. 3244–3255, 2007.
- [6] R. Sengupta, S. Rezaei *et al.*, "Cooperative collision warning systems: Concept definition and experimental implementation," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 11, no. 3, p. 143, 2007.
- [7] H. Cao, O. Wolfson, and G. Trajcevski, "Spatio-temporal data reduction with deterministic error bounds," *The VLDB Journal*, vol. 15, no. 3, pp. 211–228, 2006.
- [8] Q. Xu, T. Mak *et al.*, "Medium access control protocol design for Vehicle/Vehicle safety messages," *Vehicular Technology, IEEE Transactions on*, vol. 56, no. 2, pp. 499–518, 2007.
- [9] D. Lee, R. Attias *et al.*, "A wireless token ring protocol for intelligent transportation systems," *ITSC01*, p. 1152, 2001.