Noname manuscript No.
(will be inserted by the editor)

The Structure and Traffic Flow Anatomy of the Planet-Scale Urban Vehicular Mobility

Gautam S. Thakur · Pan Hui · Ahmed Helmy

Received: date / Accepted: date

Abstract A data-driven realistic design and evaluation of vehicular mobility has been particularly challenging due to a lack of large-scale real-world measurements in the research community. Current research methodologies rely on artificial scenarios, random connectivity, and use small and biased samples. In this paper, we perform a combined study to learn the structure and connectivity of urban streets and modeling and characterization of vehicular traffic densities on them. Our dataset is a collection of more than 222 thousand routes and 25 million vehicular mobility images from 1091 online web cameras located in six different regions of the world. Our results centered around four major observations: *i.* study shows that driving routes and visiting locations of regions demonstrate power-law distribution, indicating a planned or recently designed road infrastructure; *ii.* we represent regions by network graphs in which nodes are camera locations and edges are urban streets that connect the nodes. Such representation exhibits small world properties with short path lengths and large clustering coefficient; *iii.* traffic densities show 80% temporal correlation during several hours of a day; *iv.* modeling traffic densities against known theoretical distributions show less than 5% deviation for heavy-trailed models such as log-logistic and log-gamma distributions. We believe this work will provide a much-needed contribution to the research community for design and evaluation of future vehicular networks and smart cities.

Keywords Vehicular networks · network science · small-world · theoretical modeling · traffic analysis

1 Introduction

Research in the area of vehicular networks has increased dramatically in the recent years. With the proliferation of mobile networking technologies and their integration with the automobile industry, various forms of vehicular networks are being realized. These networks include vehicle-to-vehicle [4], vehicle-to-roadside [13], and vehicle-to-roadside-to-vehicle architectures. Realistic modeling, simulation, and informed design of such networks face several challenges, mainly due to the lack of two main factors: *i.* Underlying topology, and *ii.* Large-scale community-wide libraries of vehicular data measurement.

Topological understanding is important in accurately modeling vehicular mobility. It involves intersections, roads and their connectivity. It comes as no surprise those topological constraints like speed limits, direction, etc. impact traffic congestion, density, scenario generation, and mobility, which in turn affect the performance of any network communication protocol [2]. Thus, for accurate evaluation of a vehicular network, one should have a better knowledge of its topology.

Earlier studies in mobility modeling have clearly established a direct link between *vehicular density* distribution and the performance of vehicular networks' primitives and mechanisms, including broadcast and geocast protocols[1]. Initial efforts to capture realistic vehicular density distributions were limited by a lack of availability of sensed vehicular data[31]. Hence, there is a need to collect and conduct vehicular density modeling using larger scale and more comprehensive datasets. Furthermore, commonly used assumptions, such as Exponential distributions [1] and [30], have been used to derive many theories and conduct several analyses, the validity of which bears further investigation.

In this paper, we first study the structure and connectivity of the urban streets of six major regions and second, perform a large-scale data-driven systematic

Gautam S. Thakur
E401 CSE Building, University of Florida
P.O. Box 116120 ,Gainesville, FL 32611-6120, USA
Tel.: +1(908) 894-3095
Fax: +1(352) 392-1220
E-mail: gsthakur@cise.ufl.edu

Pan Hui
Deutsche Telekom Laboratories
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
Phone: +49 30 8353 58626
Fax: +49 391 534 78 347
E-mail: pan.hui@telekom.de

Ahmed Helmy
E426 CSE Building, University of Florida
P.O. Box 116120 ,Gainesville, FL 32611-6120, USA
Tel.: +1(352) 392-1200
Fax: +1(352) 392-1220
E-mail: helmy@cise.ufl.edu

analysis and modeling of vehicular traffic density distributions.

Recently, the departments of transportation of several regions (e.g., London, Sydney) have started to deploy traffic web-cameras to critical intersections and highways to study traffic patterns. We collected two different kind of data: *i.* geo-graphical coordinates of these locations and created a graph $G(V,E)$ as mentioned before. *ii.* To avoid the limitations of sensed vehicular data, we also utilize the existing global infrastructure of tens of thousands of video cameras providing a continuous stream of street images from dozens of regions around the world. Millions of such images captured from these available traffic web cameras are processed using a novel density estimation algorithm to build an extensive measurement dataset of spatio-temporal vehicular traffic densities.

We perform a comprehensive analysis of these data to study the structure and connectivity of urban streets and characterize the underlying statistical patterns of traffic density at individual intersections and highways of major cities. In results show that *i.* Visits to locations follow power-law distribution; *ii.* Road networks have short path lengths and large clustering coefficient, indicating small-world properties; *iii.* Temporal correlations of vehicular traffic density for individual camera locations are nearly 80% between consecutive hours, but go down to 30% for a 3-4 hours lag difference. We also investigate traffic modeling by comparing the frequencies observed in the empirical density distribution to the expected frequencies of the theoretical distribution. The result of this activity shows that the empirical values closely follow (less than 3% deviation on KS-test) heavy-tailed models such as 'Log-logistic' and 'Weibull' distributions. The contributions of this work are:

- We provide, to the best of our knowledge, by far the largest and most extensive dataset for future vehicular network analysis. This potentially addresses a severe shortage of such datasets in the community.
- We introduce a new and more practical way to look into urban street networks based on driving routes. A network graph of routes and locations depict small world properties.
- We establish heavy-tailed models such as log-logistic and log-gamma distributions as the most suitable fits for modeling vehicular traffic density.

We believe our work helps 'fill a gap' between the expected and realized necessity for the 'design and evaluation of realistic and data-driven models' for future generations of vehicular networks.

In section 2, we discuss related work, in section 3 traffic measurements and pre-processing discussed, in section 4, we discuss topological analysis of urban street maps for six different regions. In section 5, we statistically model vehicular traffic and characterize it. In section 6, we show the impact and challenges on the vehicular networks. Finally, we conclude in section 7 with future work.

2 Related Work

In this section, we discuss the related work, which is categorized in data collection and pre-processing, network analysis for urban streets and vehicular networks. In the first category, we discuss the inadequacies of existing repositories of vehicular mobility data. Next, techniques used to process image data are examined. In the past, efforts have been made to collect vehicular mobility records; by GPS traces, via loop detectors and radio sensors [14] and [16] and [34]. However, these datasets are generally not publicly available and limited in their scope, size, and geographic spread. In addition to their small timeline (typically only a few days)[1], which makes their use for longitudinal analysis limited, the methods applied were also specific to these datasets and cannot be scaled for other purposes. We believe that similar to the pedestrian trace dataset in [15], a comprehensive record of vehicular mobility is vital for a research in future vehicular networks. In contrast to the datasets described above, our dataset covers six regions, for periods of several months at hundreds of locations (see Measurement section for specifics).

Second, central to the data collection process is the image processing, designed to be computationally efficient for such a large data set. Many studies [6] have been carried out that look into aspects of both background subtraction [7] and [20] and [25] and object detection [17]. In background subtraction methods [10], difference between the current and reference frame is used to identify objects. In detection based approaches [26], learning the object features (shape, size etc.) are used to detect and classify them. In this research, we are using temporal methods for background subtraction to estimate a relative numerical value instead of counting cars. We find background subtraction is much faster, robust to outliers, applied universally and more scalable than object detection (see Measurement section for more details).

Third, attempts have been made to examine the structure and topological features of vehicular networks using applied graph measures such as centrality. In this regard, Cardillo et. al. [5] performed a 1-square mile

[1] Specifically, [1] uses 3 days of data, [14] uses traces ranging from 30 hours to a total of 400 hours from 4 to 100 cars (sensing points in that context) while [18] use a longer sample, 30 days, but only at 5 locations.

project on several cities across the world and studied local and global properties of the graphs to categorize their organic versus planned growth. They also studied the backbone of a city by deriving spanning trees based on edge betweenness and edge information. Several other studies have also found that for sustainable urban design, centrality, self-organized structures, and scaling, are driving forces [3] and [8] and [29]. In [22] and [23], the authors examined the relationship between street centrality and densities of commercial and service activities in the city of Bologna and Barcelona. We take the approach of studying the features of a region by extracting motorways that are frequently used and connected with major locations. This way, we isolate the issues of congestion and connectivity among zones of these regions.

Finally, in-car and out-car computing has shown promising results in terms of driverless cars, control interfaces for future cars that will have minimal visual demands, sophisticated traffic management systems, such as those incorporating dynamic traffic assignments [12, 9] and [19] and [32]. While further examination of these cutting edge technologies is beyond the scope of this research, we believe our approach provides a planet-scale system for data collection providing invaluable data for the development of future vehicular services and applications.

3 Measurements and pre-processing

In this section, we give details of the collected geo-location information of cameras used for the analysis of topological properties and recorded vehicular images captured from these cameras to model and characterize the vehicular traffic.

3.1 Topology Data of Camera Locations

Traffic web cameras are deployed on key intersections and highways within every city. Thus, we can assume these locations are representative of urban streets of that city. We start by recording cameras' geo-coordinates and location information to study the topological properties of urban streets. This includes latitude and longitude, zipcode, state, directional view, and camera location. Later on, in section 4.1 we use this data to create a network graph of urban streets.

3.2 Vehicular imagery data collection

There are thousands, if not millions, of outdoor cameras currently connected to the Internet, which are placed by governments, companies, conservation societies, national parks, universities, and private citizens. We view the connected global network of webcams as a highly versatile platform, enabling an untapped potential to monitor global trends or changes in the flow of the city, and providing large-scale data to realistically model vehicular, or even human mobility. Majority of these webcams are deployed by a city's Department of Transportations (DoT). Although, it's not possible to deploy them at every intersection or highway, nonetheless they are strategically placed to capture the traffic trends at critical locations. At regular intervals of time, they capture still pictures of on-going road traffic and send them in the form of feeds to the DoTs media server. We have developed crawlers that collect vehicular mobility traces from these servers. For the purpose of this study, we have also made agreements with DoTs of large regions by signing non-disclosure contracts and accepting their terms of condition to use and to collect these vehicular imagery data for several months (More information are available on individual DoT's website). We cover cities in North America, Europe, Asia, and Australia. Overall (here only six out of ten cities are presented with details in Table-I), we download 15 Gigabytes of imagery data per day from over 2700 traffic web cameras, with an overall dataset of 7.5 Terabytes containing around 125 million images. Since these cameras provide better imagery during the daytime, we limit our study to only those hours. Table 3.1 gives a high level statistics of the dataset used in this study. Each city has a different number of deployed cameras and a different interval time that captures images. We believe our study is comprehensive and reflects major trends in traffic movement. Next, we discuss the algorithm to extract traffic information from images.

3.3 Traffic Information Extraction

We aim to estimate traffic density on roads considering the number of vehicles or pedestrians crossing the road. We have a sequence of images captured by webcams. Considering our problem, we have to be able to separate information we need, e.g., number of vehicles and pedestrians from the background image, which is normally road and buildings in that image. We apply background subtraction techniques [25] and dynamic filters [10] to extract relevant traffic information. One could then use regular object detection techniques to identify and count number of vehicles in the high pass filtered image. However, this is computationally expensive and unnecessary. As an alternative, we count the number of pixels and sum their values (with a value higher than

Table 1 Global Webcam Datasets

City	# Cameras	Duration	# Images	# Routes
Connecticut	274	21/Nov/10- 20/Jan/11	7.2 million	74,801
London	181	11/Oct/10 - 22/Nov/10	1 million	32,580
Seattle	121	30/Nov/10 - 01/Mar/11	8.2 million	7,656
Sydney	67	11/Oct/10 - 05/Dec/10	2.0 million	4,422
Toronto	208	21/Nov/10 - 20/Jan/11	1.8 million	43,055
Washington	240	30/Nov/10 - 01/Mar/11	5 million	59,809
Total	**1091**	**-**	**25,2 million**	**222,323**

a certain threshold for RGB patterns that reflect darkness in the images). Due to the perspective properties of images, a vehicle will appear smaller (that is it will use less amount of pixels) when its far away from camera, whereas same vehicle may appear much bigger when in front of camera. To counter this, we exponentially weigh each pixel with increasing weights from bottom of the image to the top (where the vehicle and corresponding street width is relatively short). The cumulative sum is represented as *traffic density*. This is much faster than detecting and counting objects in an image [28]. At the same time, it is more effective, because we are looking at the percentage of the street (road), covered by vehicles (as an indicator of how crowded the street is), rather than number of vehicles. We evaluated the nature of images and found that threshold for darkness and respective spreading of headlight may cause false positive results. Since these cameras do not have night vision, we limit our study to 7am-6pm. Another threshold we calculated is based the resultant binary map obtained after the background subtraction is sent for morphological operations to remove the noise. The algorithm refine the map by removing the blobs, which have smaller area compared to the perspective properties of images. While a car count might seem preferable to a traffic density measure, there are several practical challenges. A car count requires a far greater computational cost due to the effort required to isolate each object, as current approaches are based on edge detection of objects. In addition to that, the resolution of the images is not high enough to identify vehicle's license plate number and edges in a very efficient and effective way. For privacy issues, this is actually an advantage, we cannot construct the exact trajectory of any object. Traffic congestion further complicates matters when cars occlude each other, making it difficult to segregate cars based on edge structures. In addition, vehicles at the far end of the road are small in the image and cannot be detected by these algorithms.[2] For more information, please read accompanying technical report that compare different approaches [28].

Table 2 Summary of regression analysis

Camera	df	$\beta_0(\alpha = 0.95)$	$\beta_1(\alpha = 0.95)$	R^2	p	ρ
1	100	-1.19±0.046	0.03±0.003	0.7922	0	0.91
2	100	-3.25±0.130	0.09±0.007	0.8579	0	0.92
3	100	8.16±0.045	0.10±0.005	0.9308	0	1.00
4	100	8.16±0.045	0.10±0.005	0.9308	0	1.00
5	100	8.16±0.045	0.10±0.005	0.9308	0	1.00
6	100	-2.13±0.112	0.07±0.008	0.7499	0	0.88

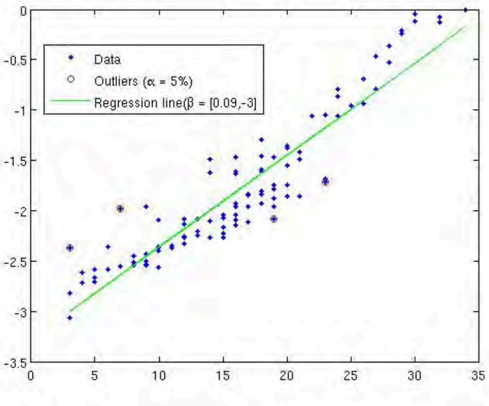


Fig. 1 A comparison of empirical traffic densities with number of cars.

3.4 Ground Truth for Validation

To test the performance of the car density capture, six cameras were selected at random and 102 images from each were examined by hand to produce a *ground truth* count for the number of cars. This ground truth was then regressed against the measured car density to check that the relationship is linear. The regression from one of the cameras is shown in Figure 1 and has a reasonable fit. There are some outliers, especially at low levels of traffic and there also appears to be a slight non-linear relationship between the ground truth and measured car density due to the warping effect of perspective (discussed above). Table 2 shows the summary

[2] Another solution could be to only count cars that are close to the camera; while this is definitely an option for video data, for snapshot data it would result in those distant cars having left the scene before the next snapshot; the net effect being that the maximum observed car count at a junction is truncated causing problems in the multivariate analysis later on.

Table 3 Parameter and details

Abbr.	Deails	Abbr.	Deails	Abbr.	Deails	Abbr.	Deails
G, $\hat{G}$	Unweighted, Weighted graph	k_m	Largest degree	d	traffic density	P	Exponential Distribution
V	Total number of nodes (camera locations)	$\langle\hat{k}\rangle$	Average weighted degree per node	L	Characteristic path length	M	Gamma Distribution
E	Total number of edges (streets)	$\hat{k_m}$	Largest weighted degree	C	Clustering coefficient	LL	Log logistic Distribution
k	Degree of a node	α	Power law exponent	L_r	Random graph characteristic path length	N	Normal Distribution
$\langle k\rangle$	Average degree per node	ρ	Correlation coefficient	C_r	Random graph clustering coefficient	W	Weibull Distribution

statistics for the regression analysis including Spearman's correlation coefficient, ρ, which seems to imply that there is a perfect non-linear correlation for camera's 3 to 5.[3] Overall, the analysis shows that while there are some errors, the relationship between the actual and measured number of cars is sufficiently clear to allow analysis at a network level.

4 Analysis of Topological Properties

In this section, we examine degree distribution and small world properties of six different regions and states in order to study the structure and connectivity of their urban street network. We represent this network by a graph $G = (V, E)$, where V is the set of camera locations as nodes and E is a set of driving segments as edges, inter-connecting the nodes of set V of the network graph G. The degree k_i of a node i in G is the number of edges incident with the node. In an undirected and unweighted G (weight = 1), the degree can be written in terms of the adjacency matrix A as

$$k_i = \sum_{j=1}^{n} A_{ij}$$

The weighted degree of each node i in undirected graph, $\hat{G}$ is $\hat{k}$, and can be written in terms of the adjacency matrix W as

$$\hat{k}_i = \sum_{j=1}^{n} W_{ij}$$

Next, we explain the graph generation process of urban street network of regions using Google Maps, and then analyze their degree distribution for unweighted and weighted cases, and finally examine their small world properties.

4.1 Network of Urban Streets

- *Segment:* A path (an edge in the network graph) that directly connects two locations.

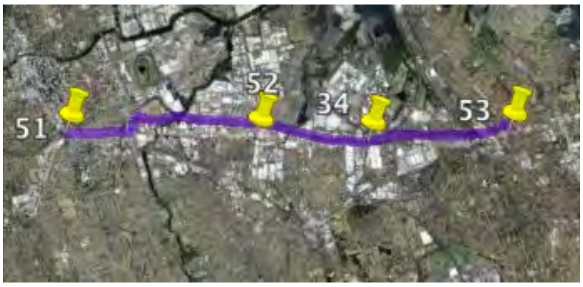


Fig. 2 An example of segment and route. The route between 51 and 53 has three segments.

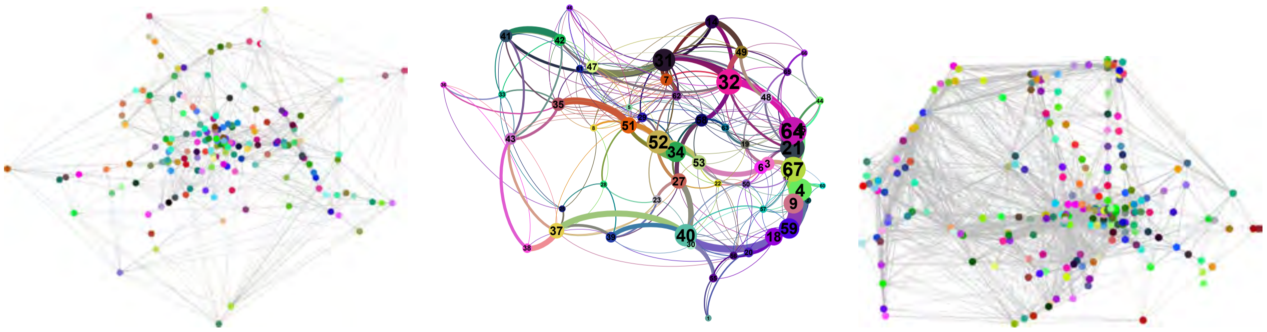
- *Route:* A path that is made up of a set of segments, when other locations appear in the order of the increasing distance from source to destination (segments corresponding to those pair of locations are added).

As mentioned before, we generate a network graph with nodes acting as camera locations and set of edges as driving route segments connecting them. Sometimes routes are made up of a set of driving segments (in case other locations are en-route between a source and a destination), as returned by Google Maps API that is shown in Figure 2. In order to generate this graph, we start by taking the geo-coordinates of a pair of camera locations and calculating the driving information between them. Next, we check for a possible subset of other camera locations that may lie en-route. All such locations are inserted in order of their occurrences and connected through intermediate segments (as edges). For example, driving from New York to San Francisco, we drive through Iowa City, Omaha, Salt Lake City, and Sacramento in that order. If no such locations exist, the source and destination are directly connected by an edge (as one big segment). We iterate this process for all pairs of camera locations, which are total $V * (V - 1)$. While doing so, we also maintain a *between* count for the traversed edges (individual segments), connecting source and destination pairs. This measure gives the frequency of a segment appearing between every pair of source and destination. We increase respective between count by one for a segment (edge) every time it is traversed. The resultant network is represented as a weighted graph showing the locations and segments (as edges) with laters' weight equal to the frequency of their appearance on multiple route.

[3] The other notation in Table 2 is standard regression notation: *df* denotes the degrees of freedom. α and β are the regression coefficients as $y = \alpha x + \beta$, R^2 is the % of variance explained, see Equation eqn:r2, p is the p-value.

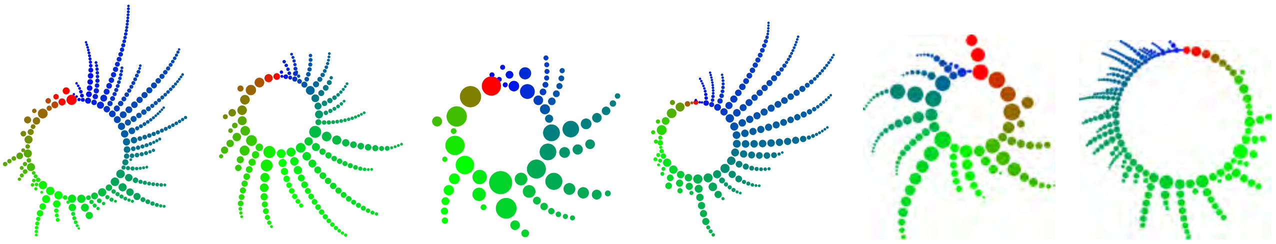
Table 4 Report of degree distribution, power-law exponent, path length, clustering coefficient, and model fitting of traffic

City	V	E	$\langle k\rangle$	k_m	$\langle\hat{k}\rangle$	$\hat{k_m}$	$\alpha(G)$	$\alpha(\hat{G})$	L	L_r	C	C_r	Dominant distribution as Best Fits (By Ranking)			Dominant distributions as Best Fits (By % Deviation KS-Test)	
													1^{st} Best Fit	2^{nd} Best Fit	3^{rd} Best Fit	≤3%	≤5%
Connecticut	274	2128	15	51	8994	32119	3.5	2.41	3.6	2.33	0.52	0.05	LL[87%]	M[11%]	P[0.5%]	LL[62%], M[15%], W[3%]	LL[94%], M[44%], W[19%]
London	181	1252	13	32	2180	7089	3.5	3.5	3.04	2.23	0.5	0.07	LL[42%]	M[39%]	W[16%]	M[34%], LL[34%], W[10%], N[0.5%]	LL[82%], M[70%], W[47%], N[7%]
Sydney	67	319	9.5	26	322	985	3.5	2.98	2.73	2.05	0.56	0.137	LL [62%]	M[32%]	N[2%]	LL[88%], M[61%], W[4%], N[2%]	LL[98%], M[88%], W[44%], N[18%]
Toronto	208	1128	10	44	7435	21323	2.8	3.5	5.02	2.5	0.6	0.05	M[46%]	W[31%]	LL[21%]	M[75%], W[58%], LL[34%]	M[94%], W[88%], LL[87%], P[4%], N[1%]
Seattle	121	513	9.8	21	1235	3376	3.5	3.5	3.3	2.27	0.56	0.087	W[36%]	LL[34%]	G[29%]	W[16%], G[14%], LL[4%]	G[55%], W[47%], LL[35%]
Washington D. C	240	3089	26.8	92	3530	15824	3.5	2.8	2.34	1.9	0.537	0.11	LL[80%]	W[11%]	G[7%]	LL[60%], W[8%], G[6.54%], E[4%]	LL[91%], W[35%], G[30%], E[14%]



(a) London (b) Sydney (c) Washington D. C

Fig. 3 A geo-laid network of urban streets of three regions is shown. The nodes are camera locations and edges are routes connecting these locations. We have shown Sydney with its weighted degree $\langle\hat{k}\rangle$ network.

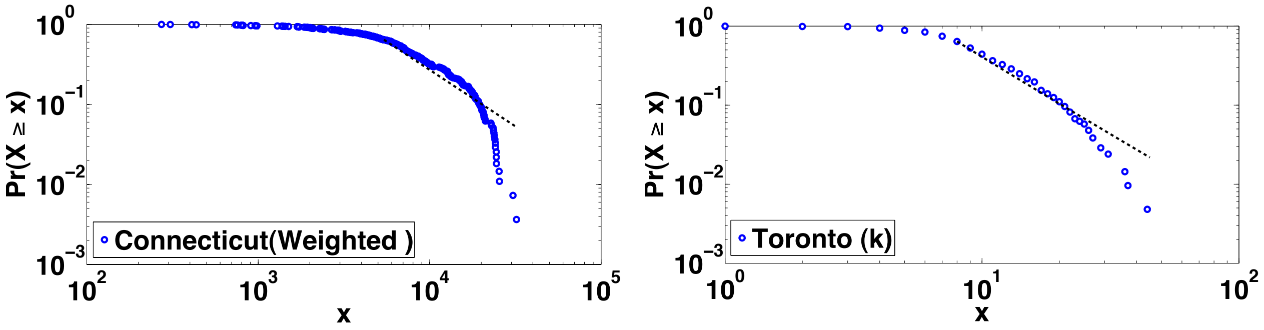


(a) Connecticut (b) London (c) Sydney (d) Toronto (e) Seattle (f) Washington D.C

Fig. 4 Radial axis layout of urban street location show a two dimensional representation of degree distribution. Each radiating axis (spar) is grouped by similar degree distribution $\langle k\rangle$. The clockwise varying of color dots from blue to red mean increase in the value of $\langle k\rangle$ for nodes. The varying sizes of dots are respective average weighted degree $\langle\hat{k}\rangle$. A larger size dots mean more weight. For Connecticut and Sydney the distribution of $\langle\hat{k}\rangle$ show power-law distributions with $2 < \alpha < 3$ and for Toronto show the same for its $\langle k\rangle$ distribution. London is an old city, with lots of small streets and intersections (hence more than one ways to reach destination), show no power-law distributions ($\alpha = 3.5$).

The most frequently taken edge segment has the largest between count. For simplicity (undirected graph analysis), we assume each street allow bi-directional traffic. In general, locations are connected by a maximum of 3-5 roadways, in our case we ease this assumption for investigating the connectivity patterns. In Fig.3(b), we show the example of a weighted graph of Sydney generated with 67 camera locations. The underlying process of generating this network graph is computationally expensive [27], nonetheless it has many benefits: *i.* We use Google Maps API to calculate all possible routes and intersections, today, anyone planning to travel, accesses maps via Google or like services. *ii.* We are assured that resultant graph filters-out non-frequent routes, which help to better explain the cause of traffic congestion on frequently taken routes and locations. *iii.* The recommendations can be made to generate dynamic routes from diverting the traffic on already congested segments. There are several variations of *between* count.

- *Unweighted Graph*: We baseline the *between* score of a street (edge) to one if it has ever appeared.
- *Weighted graph by distance:* The weights on the edges can be replaced by actual driving distance. Thus, recommendation can be made in case shortest path is available between a pair of source and destination.
- *Weighted graph by distance and between score*: The weights on the edges are a combination of distance and *between* score. It helps to discover overhead and congested segments in the network.



(a) Connecticut ($\langle \hat{k} \rangle$, $\alpha = 2.41$)

(b) Toronto ($\langle k \rangle$, $\alpha = 2.8$)

Fig. 5 A CDF $P(x)$ and its maximum likelihood power-law fits for two locations.

4.2 Analysis of Degree Distribution: Unweighted Case

We study the number of connections that camera locations have with one another. It helps to analyze the connectivity and probability of taking alternate routes to the destinations. A network analysis of Toronto's cameras shows their degree distribution is follow power-law distributions. As evident in Fig.4, the radial axis graph of Toronto clearly shows only 2-3 locations with large degree distribution. In Fig.5, an exponent value of $\alpha = 2.8$ shows that the maximum likelihood power-law fit for the degrees of its few locations have a long right-side tail of values that are above the mean. A value x obeys a power-law if it is drawn from a probability distribution p as:

$$p(x) \propto x^{-\alpha}$$

α a constant known as exponent parameter. The usual value of exponent lies in the range $2 < \alpha < 3$ with some exceptions.

Above results indicate that such locations have much higher connectivity with rest of the one-hop far locations. On the other hand, if they are removed from the network, average path length will increase, and location pairs will become disconnected and traveling between them will become impossible.

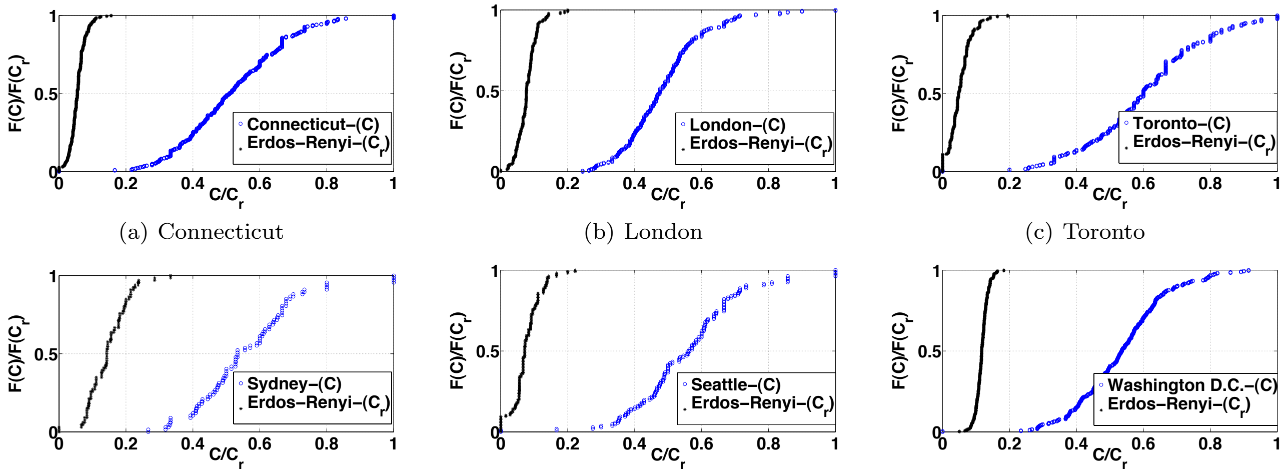
4.3 Analysis of Degree Distribution: Weighted Case

The weighted degree of a camera location is calculated based on the frequency of its connected edges that have appeared between any pair of source and destination. Using Google Maps, we have calculated shortest path between all pairs of locations, and the list of locations that are on en route. Therefore, it is possible that few locations have been traversed more often than other, making them the most visited locations. In our study, we find the locations belonging to Sydney and Connecticut demonstrate a power-law distributions, which means they create an hour glass model, making most of the traffic to pass through few locations. It also makes them susceptible to traffic congestion and closures. In Fig.4, we see the distribution of node sizes representing weighted degrees for Connecticut and Sydney, with power-law exponent $\alpha = 2.41$ and 2.98 respectively in Table-4. In Fig.5, a cumulative distribution function for maximum likelihood fit for Connecticut and Toronto is shown.

Thus, while Toronto is skewed on connectivity, Connecticut and Sydney are skewed on visiting same locations again and again. We can say that traffic congestion in Toronto appears because of geometry of locations, while for Connecticut and Sydney its because specific routes have been traversed. The city of London appears to have even distribution for both metrics, as evident in its radial layout in Fig.4 and Table-4. We can say that London network is more resilient than other regions, with lot of small and inter-connecting streets, exhibiting properties of an historic city's growth.

4.4 Small World Analysis

We investigate that network of urban streets of all six regions clearly exhibit small world properties. In general, a network with small world should have small average path length ($L < 6$) and large clustering coefficient ($0.4 < C < 1$). We make a basis for a fair comparison, by using Erdos-Renyi G(n,M) [11] model to generate a random graph for each city separately, with $n = V$ and $M = E$. To ascertain our structure, we examine C against C_r for each city - for C to be extreme in that distribution and greater than the ninety-fifth percentile. Next, we calculate the average path length (L) and clustering coefficient (C) of the six regions' networks and compared them against L_r and C_r of random graphs respectively, as shown in Table-4. We find that networks of all six regions have small average



(a) Connecticut (b) London (c) Toronto

(d) Sydney (e) Seattle (f) Washington D.C.

Fig. 6 A CDF of C and C_r show large values of clustering coefficient for regions network to random graphs, indicating priors' network structure exhibiting small world properties.

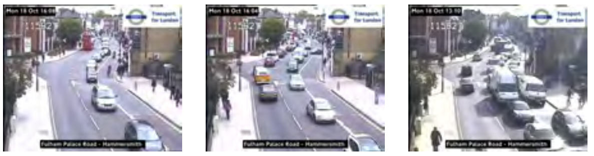
path length ($\forall L < 6$) and large clustering coefficient ($\forall C \rightarrow 1$, $C \gg C_r$), with Toronto having largest value of clustering, $C = 0.6$. The CDFs of clustering coefficient are shown in Fig. 6 - *i.* a quick convergence for the random graph, indicating very small clustering coefficient values of C_r *ii.* All values of $C > 0.3$, and large gaps in curves indicating network of regions exhibiting strong small world properties.

5 Traffic Modeling and Characterization

We studied connectivity of urban streets, now we turn to model and characterize the traffic density on these streets. We will see, how the traffic is correlated with itself for several hours of the day. Later, we will use known theoretical distributions to model traffic densities.

5.1 Traffic Flow Auto-Correlation

We investigate correlation coefficients (ρ) to measure the degree to which traffic from a camera is linearly associated with itself for 42 days. In our case, we are using this to analyze the change in traffic densities. We analyze the correlations for 1-4 hour lags for each camera against itself during 12 hours of the day, from 7 AM to 6 PM. For example, we investigate what the correlation is between the traffic at 7 AM and 8 AM (1-hour lag), 1 PM and 3 PM (2-hour lag) etc. In Fig.7, we show CDF for various hours lag of the day. For the city of Sydney the hourly traffic change is highly correlated, almost 80% of cameras' next hour traffic is 70% correlated to its current hour. For next two hours from the current, the traffic for 80% of the cameras are only 50% or less correlated. And around 60% cameras have only 30% correlation for a time lag of 3-4 hours. While in case of the city of London, the next hour traffic density for 80% cameras is close to 60% correlated to the current hour. It goes further down to 30% for next two hours and around 15-20% for a 3-4 hour difference. *Thus, vehicular traffic has temporal richness, which in-turn affects the mobility of vehicles and therefore, have an impact on the performance of routing protocols [2].* Similar trends are observed in other regions, but omitted here for brevity.



(a) $d = 2023, 0.28$ (b) $d = 5400, 0.55$ (c) $d = 9230, 0.93$

Fig. 8 Traffic with varying densities[(a)low/(b)medium/(c)high] is shown. The first value is the result of background subtraction and later is the normalized value.

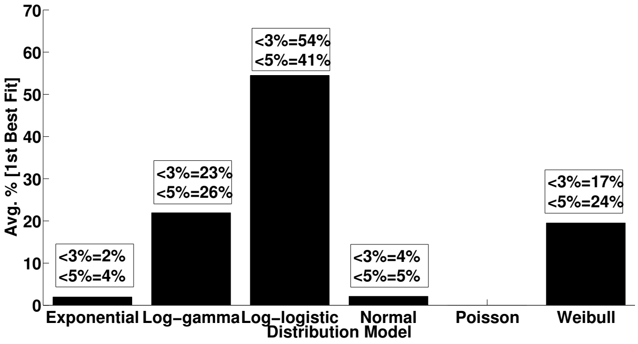
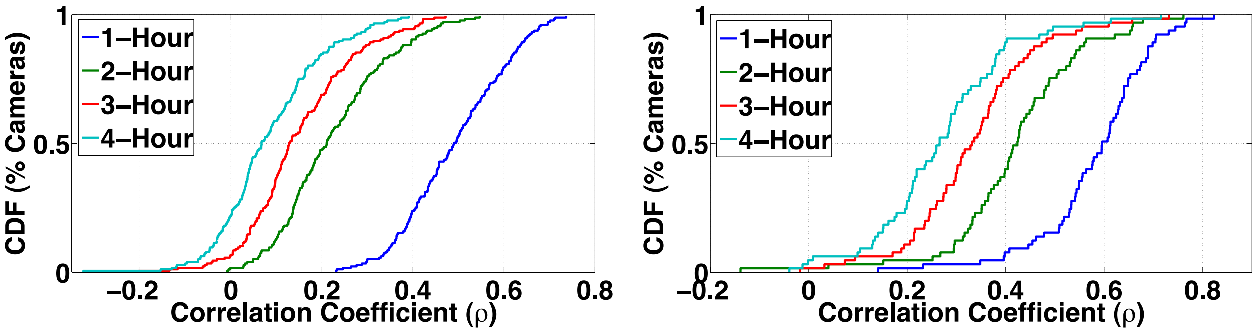


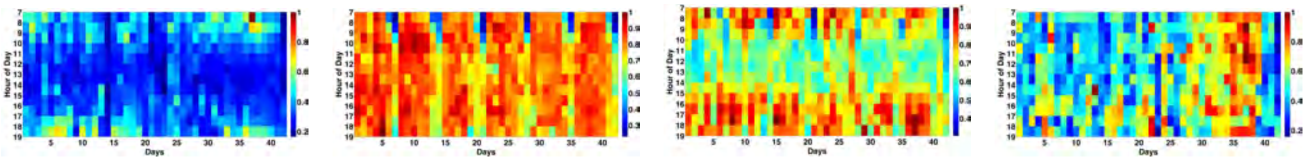
Fig. 10 Best fits for six regions. The values in the box show deviation.



(a) London

(b) Sydney

Fig. 7 CDF showing correlation of traffic densities between hour differences of the day.



(a) Low Traffic

(b) High Traffic

(c) Periodic Traffic

(d) Random Traffic

Fig. 9 Several variations in traffic densities across 42 days traffic monitoring are shown. Fig-(a) show relatively mild traffic during various hours of the day, while (b) show high traffic recording for the full trace periods. In Fig-(c) we find a regularity patterns during the morning and evening hours when the traffic is relatively higher to afternoon hours. A random traffic pattern is recorded in the last.

5.2 Traffic Modeling and Characterization

Here, we focus on modeling the arrival process of traffic (traffic density value) in equal intervals of times against known theoretical distributions. In Fig.8, we show three traffic scenarios of varying intensities from low to fully congested location, captured by the density parameter(d).The objective of this study is to help understand the underlying statistical patterns. We already filtered these for the purpose of showing the maturity in our studies to select and identify the statistical patterns without much deviations. To ensure the validity, we also performed several goodness of fit test using Maximum likelihood estimation (MLE) and Kolmogorov-Smirnov test to measure average deviation and compare the values in the density vector to known distribution. We systematically model individual locations' empirical traffic density distribution against well known theoretical ones. In Fig.9, we show four different locations with changing traffic densities during 12 hours for 42 days. *This result invalidates a general notion of 'rush hours' that traffic is relatively higher only during morning and early evening hours.* In order to match, we use five theoretical distributions: Exponential, Gamma, Log-Logistic, Normal and Weibull. We find that traffic at individual cameras can vary a lot, but in general log-logistic, Gamma and Weibull distributions can capture some of the key features. We rank these distributions (based on KS-tests) in Table-3, with four out of six regions' individual locations have log-logistic as the 1^{st} best fit, while Toronto has Gamma distributions. In Table-3, we show dominant distributions at 3% and 5% deviation using the KS-test. In Fig.10, results show the dominance of distributions for all the locations from all six regions. Overall, the empirical data closely matches log-logistic and Gamma distributions. We find that even on regions' aggregate traffic levels, the log-logistic distributions provide a good estimate of empirical data. These results are realistic scenarios, and can be used as input for simulators to evaluate the performance of vehicular routing protocols.

6 Future Application to Vehicular Networks

The experience gained from the analysis and modeling of traffic densities potentially aids in future design and evaluation of vehicular networks. Today, most of the simulation tools input generic or random scenarios and disregard the challenges brought by mobility in vehicular networks [2] and [24] and [33]. In our case, the benefit of urban street analysis and large dataset of realistic traces, and its modeling results prove to be very helpful in developing rich scenarios for testing protocols, network dynamics, scalability of traffic, topology size estimation, and the analysis of traffic patterns. The data-driven realistic simulation tools and mobility models are necessary for accurate evaluation of vehicular routing protocols and services. However, our analysis shows that traffic characterization and communication network analysis tools (e.g., ns2) are separately developed and therefore lack a tight integration [24] and [21]. Our gathering and analyzing real traffic data can aid in identifying metrics (e.g., spatio-temporal density) to develop data driven mobility models and simulators. The unique challenges (e.g., high speed, intermittent connectivity) in inter-vehicle [4] and car-to-roadside [13] communication require the development of robust and efficient routing protocols. We can use the cameras' geo-coordinates and their traffic density distribution to develop and test new performance metrics and protocols. In the future, we aim to focus on developing realistic and data-driven models. We have also plan to make this dataset available to the research community and extend our existing work to study centrality measure for all the cities.

7 Conclusion

We know topological properties (like directions and lanes) impact the movement of vehicular traffic on roads. In this paper, first we have discussed an approach to create a network of urban streets from driving directions and second use of vehicular imagery snapshot images from freely available online cameras for traffic analysis. Our results have shown that for three regions (Connecticut, Sydney, and Toronto), during several trips, visits to their locations and streets exhibit a power-law distributions. A temporal auto-correlation of 80% is evident for traffic densities in those three cities for consecutive hours (1-2 hours) of the day. In London, high and variable traffic pattern. We have observed a stable periodicity of traffic density for many days (42 days) corresponding to weekdays and weekends. This is an important result, and can aid in developing futuristic traffic prediction models. We have also found that empirical traffic densities closely follow (with less than 3% deviation) theoretical distributions like Log-logistic and Weibull. We believe our work will provide much needed contribution to the research community.

References

1. Bai, F., Krishnamachari, B.: Spatio temporal variations of vehicle traffic in vanets: facts and implications. In: Vanet (2009)
2. Bai, F., Sadagopan, N., Helmy, A.: The important framework for analyzing the impact of mobility on performance of routing protocols for adhoc networks. Ad Hoc Networks **1**(4), 383–403 (2003)
3. Batty, M.: Network geography: Relations, interactions, scaling and spatial processes in gis. RGIS (2003)
4. Blum, J., et. al.: Challenges of intervehicle ad hoc networks. ITS, IEEE Tran. on (2004)
5. Cardillo, A., Scellato, S., Latora, V., Porta, S.: Structural properties of planar graphs of urban street patterns. PRE **73** (2006)
6. Chandler, R.E., et. al.: Traffic dynamics: Studies in car following. Jour. of OR (1958)
7. Cheung, S.C.S., Kamath, C.: Robust background subtraction with foreground validation for urban traffic video. EURASIP **2005**, 2330–2340 (2005)
8. Crucitti, P., Latora, V., Porta, S.: Centrality measures in spatial networks of urban streets. PRE (2006)
9. Dave, N.K., Vaghela, V.B.: Vehicular Traffic Control: A Ubiquitous Computing Approach (2009)
10. Elgammal, A., et. al.: Non-parametric model for background subtraction. In: Computer Vision, LNCS (2000)
11. Erdos, P., Renyi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**, 17–61 (1960)
12. Gary, B., Porter, J.M.: Ubiquitous computing within cars:designing controls for nonvisual use. IJHCS (2001)
13. Jiru, J., Eilers, D.: Car to roadside communication using ieee 802.11p technology. Industrial Ethernet Book (2010)
14. Kaul, S., Gruteser, M., Rai, V., Kenney, J.: On predicting and compressing vehicular GPS traces. In: ICC (2010)
15. Kotz, D., Henderson, T.: CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. IEEE Pervasive Computing (2005)
16. Krumm, J.: Trajectory analysis for driving. In: Computing with Spatial Trajectories. Springer (2011)
17. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: ICIP (2002)
18. Meng, Q., Khoo, H.L.: Self-similar characteristics of vehicle arrival pattern on highways. Jour. of Transportation Engineering **135**(11), 864–872 (2009)
19. Moore, M.M., Lu, B.: Autonomous Vehicles for Personal Transport. SSRN eLibrary (2011)
20. Piccardi, M.: Background subtraction techniques: a review. In: Systems, Man and Cybernetics, IEEE International Conference on (2004)
21. Piórkowski, M., et. al.: Trans: realistic joint traffic and network simulator for vanets. Sigmobile CCR (2008)
22. Porta, S., Latora, V., Wang, F., Rueda, S., Strano, E., Scellato, S., Cardillo, A., Belli, E., Crdenas, F., Cormenzana, B., Latora, L.: Street centrality and location of economic activities in barcelona. Urban Studies (2011)
23. Porta, S., Latora, V., Wang, F., Strano, E., Cardillo, A., Scellato, S., Iacoviello, V., Messora, R.: Street centrality and densities of retail and services in bologna. EPD (2009)
24. Stanica, R., Chaput, E., Beylot, A.: Simulation of vehicular ad-hoc networks: Challenges, review of tools and recommendations. Computer Networks (2011)
25. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE CVPR (1999)
26. Sun, Z., et. al.: On-road vehicle detection: A review. IEEE Tran. on PAMI (2006)
27. Thakur, G.S.: Generating Urban Street Network using Google Maps API (2012). URL http://www.cise.ufl.edu/∼gsthakur/globalsense.shtml
28. Thakur, G.S., Ali, M., Hui, P., Helmy, A.: Comparing background subtraction algorithms and method of car counting. CoRR (2012)
29. Wilson, A.: Complex spatial systems: modeling foundations of urban and regional analysis. PH (2000)
30. Wisitpongphan, N., et. al.: Routing in sparse vehicular ad hoc wireless networks. IEEE Comm. (2007)
31. Yeo, J.: Crawdad: a community resource for archiving wireless data at dartmouth. Sigcomm CCR (2006)
32. Yoshihiro, S., Minoru, K., Yukio, K.: Driverless car traveling guide system. USPT (1989)
33. Yousefi, S., et. al.: Vehicular ad hoc networks (vanets): Challenges and perspectives (2006)
34. Zhang, D., Li, N., Zhou, Z., Chen, C., Sun, L., Li, S.: iBAT: detecting anomalous taxi trajectories from GPS traces. In: Ubicomp (2011)