Data-driven Study of Urban Infrastructure to Enable City-wide Ubiquitous Computing

Gautam S. Thakur
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
thakurg@ornl.gov

Pan Hui
Hong Kong University of Science and Technology and Telekom Innovation Laboratories, Berlin
pan.hui@telekom.de

Ahmed Helmy
CISE, University of Florida
Gainesville, Florida, USA
helmy@cise.ufl.edu

ABSTRACT

Engineering a city-wide ubiquitous computing system requires a comprehensive understanding of urban infrastructure including physical motorways, vehicular traffic, and human activities. Many world cities were built at different time periods and with different purposes that resulted in diversified structures and characteristics, which have to be carefully considered while designing ubiquitous computing facilities. In this paper, we propose a novel technique to study global urban infrastructure, with enabling city-wide ubiquitous computing as the aim, using a massive data-driven network of planet-scale online web-cameras and a location-based online social network service, Foursquare. Our approach examines six metropolitan regions' infrastructure that includes more than 800 locations, 25 million vehicular mobility records, 220k routes, and two million Foursquare check-ins. We evaluate the spatio-temporal correlation in traffic patterns, examine the structure and connectivity in regions, and study the impact of human mobility on vehicular traffic to gain insight for enabling city-wide ubiquitous computing.

Keywords

Urban Infrastructure, Vehicular Traffic, Network Science

Categories and Subject Descriptors

H.2.8 [**[Database Management]**]: Database Applications-Data mining, Image databases, Spatial databases and GIS.

General Terms

Experimentation, Human Factors, Measurement.

1. INTRODUCTION

Ubiquitous computing is deemed vital for the development of environment friendly and sustainable smart cities. It's use has been realized for example, in electronic road payment systems, computer-driven mass transits, smart-postal, and mobile networks [1, 11]. On the other hand, the urban infrastructure such as geographical spread and road networks will influence the design and deployment of such ubiquitous computing systems on urban scale [11, 17]. Hence, this tight coupling calls for a thorough understanding of urban infrastructure for the realization of city-wide ubiquitous computing effort. There are several factors that impact the understanding of urban infrastructure. In general, the topological features to study the infrastructure of an urban setting primarily involves it's geographical spread and area, network of motorways, structures (such as buildings and dams) and human population density. In this regard, many studies have examined a stark difference among self-organized cities that are evolved because of some historical processes versus those that are the result of a single-plan, mostly producing a grid-like structure. The historical cities have observably more densely packed network of intersections and small motorways, less fragmented and decentralized geographic expansion [3, 4, 16]. They are also more populated, inhabited, and demonstrate more complex and dynamic eco-systems. On the other hand, man-made designed cities are more structured with evenly distributed spatial flows and sparse landscaping. Other meta-physical factors such as globalization, socio-economic and financial viability, technological advancement, and politics also affect our understanding of urban infrastructure. Moreover, these diversities bring a principle challenge to design and develop practical tools [5] for measuring and quantifying their interacting effects, popularly known as emergent properties. In addition to this, the evaluation criterion should be generic enough to apply to any setting (urban infrastructure). Essentially, a study of these activities will provide a significant insight for enabling city-wide ubiquitous computing environment. For instance impact of aforementioned features are well documented for Singapore and Korea in [1]. Several other studies have also shown idiosyncrasies in deployed systems based on the structure and function of urban infrastructure [8, 7].

Recently, Department of Transportation (DOTs) across several metropolitan areas have started to deploy traffic web-cameras. These cameras are strategically located and positioned towards motorways to enable the monitoring of vehicular traffic. At a constant interval, they capture snap-shots of traffic conditions, which are then available for viewing on DOTs' media servers. We have collected and processed more than 25 million such images to generate longitudinal time series dataset of traffic densities for more than 800 locations spread in six regions (Connecticut, London, Seattle, Sydney, Toronto, and Washington D.C.) around the world. In this paper, we harness the power of these cameras and use this dataset to study vehicular traffic conditions and use cameras' geo-graphical spread to analyze the topological properties of these regions. In the current scenario, Online Social Networks (OSNs) such as Facebook and FourSquare are tightly integrated with our life-style. They provide numerous ways to share our diurnal patterns, presence and movements with the rest of the world. In order to study the hu-

man dynamics in these six regions, we have collected anonymous spatio-temporal footprints of human activities through FourSquare. In this work, we integrate these two different datasets (Vehicular and Human) and use generic tools such as network centrality to comprehensively study and reason the urban infrastructure of these six regions. By the way, these regions are a mix of planned man-made cities and self-organized historical cities, as discussed before.

In our approach, we first examine the nature of vehicular traffic. In that we study traffic patterns (regular or random), evaluate traffic correlations across all location pairs and their stability (auto-correlation) in their region. We also locate hotspots (congestion prone zones) and similar traffic locations in individual regions. Second, we examine the spatial features of these locations. In that we correlate travel distance and time in these locations and reason about reachability. Then we turn the geographical map of these regions into network graphs. We employ various centrality measures [12] in order to access the structure and connectivity that have a big impact on the behavior (topology) of the system. Finally, we examine the human dynamics with vehicular traffic on these locations and reason and study their correlation in urban settings. To summarize, our contributions are:

- We propose a novel data-driven technique to use global infrastructure of traffic cameras to perform a longitudinal study of urban infrastructure.
- We provide a comprehensive study into the topological features by integrating diversified data of vehicular traffic, urban streets, and human dynamics. In future, we plan to release this dataset to the research community.
- Our approach has involved the use of generic and systematic techniques that can be scaled and used in any settings.

The rest of the paper is organized as follows: Section 2 has details of dataset and processing techniques. In Section 3, we examine traffic patterns and in Section 4, we perform network evaluation of urban street maps. In Section 5, we study human dynamics with vehicular density patterns and finally Section 6 concludes our work.

2. MEASUREMENT AND VEHICULAR DENSITY ESTIMATION

Table 1 summarizes the dataset used in this research; six regions/cities, the time span of the samples, the sampling rate and the number of camera's/sample locations. On average, we download 15 gigabytes of imagery data per day from over 2,700 traffic cameras, with an overall dataset of 7.5 terabyte containing around 125 million images. In this paper, we have selected six regions with similar time granularity of traffic snap shots, as shown in Table 1. The subset of dataset used has 25M records. Figure 1 shows the distributed system architecture for vehicular imagery collection on planet-scale that we have built at Deutsche Telekom Labs, Berlin. Figure 2 shows a geological snapshot of the cameras deployed in London and Sydney, as an example. The area covered by the cameras in London is 950km^2 while that in Sydney is 1500km^2. Finally, note that since these cameras do not have night vision, we limit our study to the hours between 7am and 6pm.

2.1 Background Subtraction

The snapshots taken by every traffic camera (at intervals ranging from 20-60 seconds) first pass a background estimation and subtraction phase. These are then used to estimate the *traffic density*

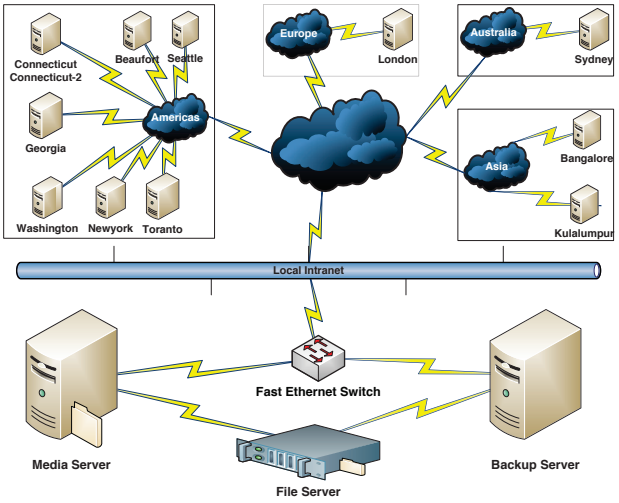
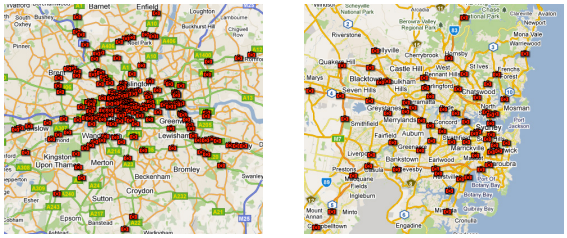


Figure 1: Distributed system architecture for vehicular imagery collection on planet-scale.



(a) London (b) Sydney

Figure 2: Traffic cameras in London and Sydney. The red dots show the location of cameras deployed giving an idea of their distribution in the city.

arriving per unit time as opposed to a car count. 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. 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.[1]

Background subtraction is a standard method for object localization in image sequences with fixed cameras, where the frame rate is lower than the velocity of the objects to be tracked (i.e. cars move out of the scene typically at a rate exceeding 1 minute). The models of background are based on the observation that *background* does not change significantly (in comparison to foreground/objects) across time. Any part of an image that does not fit with that model is deemed as *foreground/object*. These foreground regions are then further processed for the detection of desired objects. The background model used here assumes that the distribution of background pixel values may be modeled as a weighted sum of Gaussian distributions. Our approach follows closely to those proposed by [2, 13, 14] because of their reliability and robustness to sensitive changes

[1]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 1: Global Webcam Dataset

Region	# of Cameras	Duration	Interval	Records	Database Size	Routes
Connecticut	120	21/Nov/10- 20/Jan/11	20 sec.	7.2 million	435 GB	74,801
London	182	11/Oct/10 - 22/Nov/10	60 sec.	1 million	201 GB	32,580
Seattle	121	30/Nov/10 - 01/Mar/11	60 sec.	8.2 million	600 GB	7,656
Sydney	67	11/Oct/10 - 05/Dec/10	30 sec.	2.0 million	350 GB	4,422
Toronto	89	21/Nov/10 - 20/Jan/11	30 sec.	1.8 million	325 GB	43,055
Washington	240	30/Nov/10 - 01/Mar/11	60 sec.	5 million	400 GB	59,809
Total	**819**	**-**	**-**	**25.2 million**	**2311 GB**	**222323**

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

in the lighting conditions. In our approach, the observed pixel value is modeled by a weighted sum of Gaussian kernels. Let x_t represent a pixel value in the t^{th} frame, then the probability of observing this value is assumed to be:

$$p(x_t) = \sum_{i=1}^{K} w_i^t * \mathcal{N}(\mu_{i,t}, \Sigma_{i,t}) \quad (1)$$

where $\mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$ is the i^{th} kernel with mean $\mu_{i,t}$ and covariance matrix $\Sigma_{i,t}$, and w_i^t is the weight applied to that kernel such that $\sum_i w_i^t = 1$. We assume that RGB channels are uncorrelated thus the covariance matrix for each kernel is diagonal.[2] When a new frame arrives, the pixel values are compared to the kernels to determine if it is likely that this value was drawn from a distribution with $\mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$ (using for example a 95% confidence interval). If so, $\mu_{i,t}$, $\Sigma_{i,t}$ and w_i are updated using exponential filters; if not, a new kernel is created and the existing kernel with the lowest w_i is eliminated (see [14] for specifics). Short lived kernels and their associated pixels are deemed to be possibly foreground producing a binary map. Morphological operations are then applied to this map to remove noise and any blobs with area smaller than a certain threshold. The view of most cameras used in this study is along the direction of the road and this perspective skews the size of objects on an image [6]. To counter this effect, we weigh each foreground pixel with the exponent of it's distance from the bottom of the image. Thus a pixel in the bottom of the image will be weighted less (objects appear larger at the bottom than on the top) than a pixel at the top. While this weighting is not exact and does produce some warping as we shall see in the ground truth validation section; the warping is not excessive given the advantage that weighting is simple and does not require manually tuning at each camera.

2.2 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 if the relationship is linear. The regression from one camera is shown in Figure 3 and shows 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 statistics for the

[2]Thus reducing the number of unknown parameters.

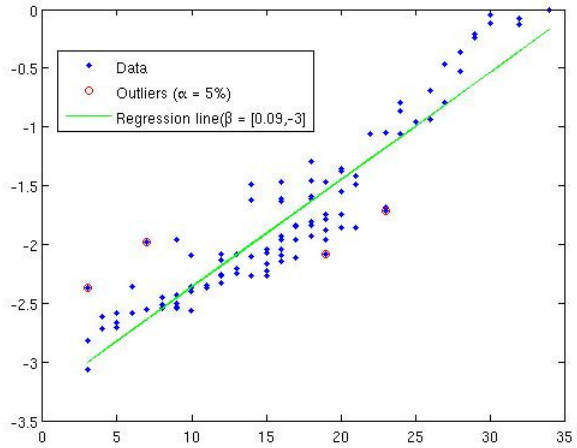


Figure 3: A comparison of traffic densities with number of cars.

regression analysis including Spearman's correlation coefficient, ρ, which seems to imply that there is a perfect monotonic 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 statistically accurate.

3. SPATIO-TEMPORAL ANALYSIS

In this section, the characteristics of the traffic data are analyzed across time and location; spatio-temporal analysis. To begin, Figure 4 shows the average density for all cameras for two cities; Sydney and London. As can be seen there is an expected diurnal pattern; a morning and evening rush hour. However, the time series also exhibits a high variance as can be seen by the 95% confidence intervals (dashed line). This underlies the fact that while a strong diurnal pattern is evident on average days, this may not be the case for some particular days. Comparing the two cities, it is interesting to note that Sydney has a significantly lower average than London but a higher variance. This is contrary to typical time series where a higher mean is usually accompanied by a higher variance. The most likely explanation is that in a city with high congestion there is little room for maneuver; the city is quite simply congested every day and so the traffic density every day looks broadly similar. The implication of this behavior is that with future efforts to relieve congestion, comes increased difficulty in predicting congestion.

The next step examines the daily patterns in the average density for a city to see if the large variabilty observed in Figure 4 can be

[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.

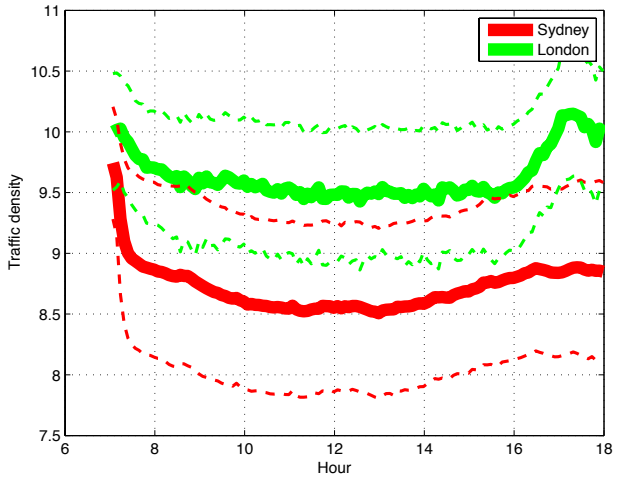
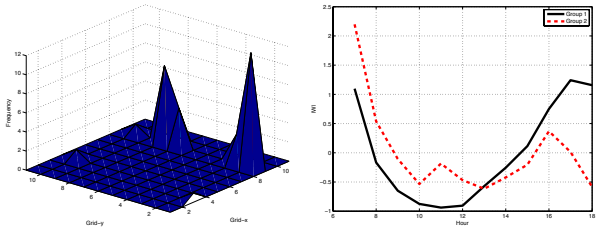


Figure 4: Average density for regions of Sydney and London



(a) Kohonen map (b) Day-types

Figure 5: Kohonen results: (a) Kohonen map for Sydney showing 2 days (b) the 2 patterns associated with two peaks in (a).

explained. The data is decomposed into a 30 $\times$ 12 matrix of 30 daily patterns, each 12 hours long. A Kohonen neural network is then used to classify these 30 daily patterns into groups called *day-types*. A Kohonen neural network is composed of a grid of output points (in this case a 2-D grid) where each grid point has an associated pattern; the patterns in adjacent grid points being similar. When a pattern is presented as input to the network it is compared to the grid patterns and the closest match is declared the winner. The training algorithm consists of beginning with random patterns on the grid points and adjusting the (at first random) winner and its neighbors until the data has been sifted into its constituent groups. The specific algorithm used here is explained in detail in [9]. Figure 5(a) shows the resulting map constructed from the Sydney data set. As can be seen, there are two very distinct day-types in the data covering approximately half the data each. The corresponding patterns at those two grid locations are also shown in Figure 5(b); these are the *archetypal* day-types. These show an intriguing result; for day-type one (solid black), the evening rush is roughly the same as the morning rush with the expected lull in the middle; for the second day-type however, the morning rush is dominant with a larger peak than expected during the afternoon and an evening peak that is much less than the morning rush. The second day-type can be partially explained by the weekends but not completely so (as it accounts for almost half the data), thus the traffic in this data set is not as predictable as originally may have been assumed. For traffic management it is obviously important to know the different day-types that exist in the network and when they are likely to occur.

In general, it has been observed that vehicular traffic has certain pattern and show periodicity in nature. In order to enable ubiquitous computing, it is important to study and quantify them. Particularly, we ask following questions:

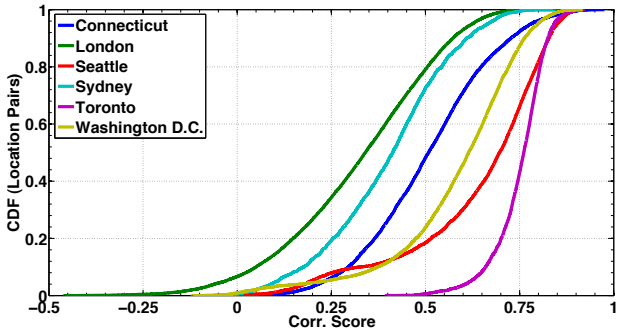


Figure 7: A CDF showing the distribution of traffic densities that are correlated across the locations.

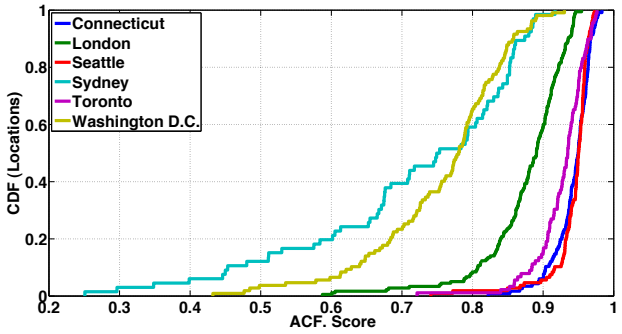


Figure 8: A CDF showing the distribution of average autocorrelation (weekdays) for locations of six cities.

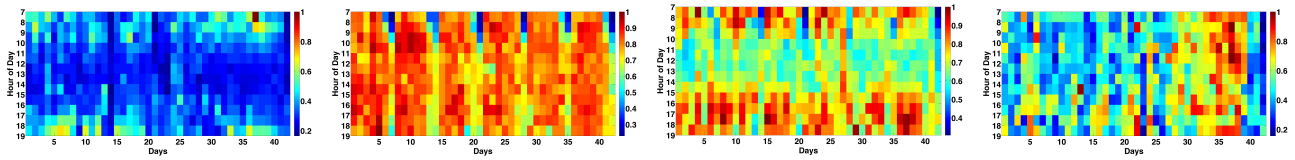
Q.1: *What does the traffic distribution look like across several hours for multiple days?*

Q.2: *How is the traffic distributed across several locations of a region?*

Q.3: *How is the traffic correlated with itself? Is the traffic predictable?*

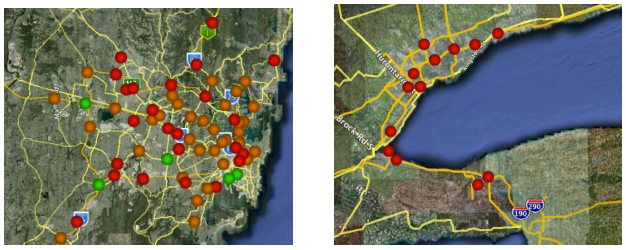
In order to answer the first question, we hourly sample the traffic dataset for a period of 42 days and study their density distribution patterns. The results indicate that cameras have varying traffic distribution against the popular notion of 'rush hours'. In Figure 6, we show the traffic density distribution from four sampled cameras. It is evident that Figure 6(a) has very low traffic throughout the 12-hours period for all 42 days. In Figure 6(b), consistently high traffic is recorded for a street in London, with relatively less traffic during the weekends (day 7, 14, 21 on weekends). We also find the periodicity in traffic during the morning and evening hours in case of Sydney, as shown in Figure 6(c). Thus, here the temporal activity reaches its maximum value during the morning and evening hours while it is low during the afternoon hours. Finally, some random patterns are observed in Figure 6(d). In general these results reject the notion of one-size-fits-all and provide essential input in deploying the ubiquitous system that conforms to periodicity.

To answer the second question, we perform correlation analysis of traffic time series for all pairs of locations of a region. Our results in Figure 7 indicate that traffic distribution across 50% of Toronto's locations is 75% correlated and 60% of Seattle's location



(a) Low Traffic Density (b) High Traffic Density (c) Periodic Traffic Density (d) Random Traffic Density

Figure 6: Several variations in traffic densities across six-weeks traffic monitoring is 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 than afternoon intervals. A random traffic characterization is recorded in the last.



(a) Sydney (b) Toronto

Figure 9: (a) Sydney traffic similarity. (b) Hot-spots in Toronto.

is 50% correlated. It make sense that the correlations are high, since many cameras that are deployed in these two regions are on highways, which generate consistent traffic patterns. In case of Sydney and London, we find that deployed cameras are within city limits (business places and residential area) and believe to have uncorrelated traffic distribution patters. These results provide an important insight into the categorization of various motorways based on the distribution of traffic that is correlated to each other. In Figure 9, we point out similar traffic patterns (same color bulbs) for Sydney and high intensity traffic (hotspots) locations in Toronto.

Next, to answer the third question, we sample the traffic of each location and calculate auto-correlation function to examine the variability in the patterns across several weekdays. The result of this analysis is shown in the Figure 8. We find that for Seattle, Toronto, and Connecticut the traffic is highly auto-correlated. While for London, we have registered some variation during the weekdays, and the least auto-correlated are Sydney and Washington D.C., where the traffic is nearly 70% autocorrelated with 40% of their individual locations (region wise).

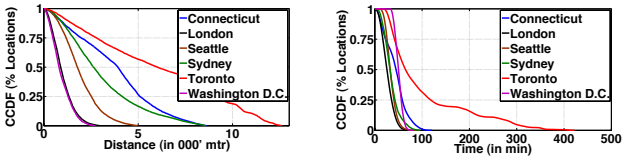
These findings are very interesting for Seattle and Toronto, where the distribution of traffic is not only correlated among its locations, but also highly correlated with itself for individual locations (more predictable). While Sydney and London demonstrate lot of variance in their auto-correlations and correlations across several locations. Overall, we believe, this study will provide lot of insight into the deployment of ubiquitous systems such as self-driving vehicles and transit systems.

4. NETWORK ANALYSIS

In this section, we perform spatial analysis of motorways and intersections to study geographical spread, connectivity, and city dynamics. The details of the routes used in this study are listed in Table 1. For more information regarding route calculation, please refer [15]. First, we examine travel distance and time among all locations of individual regions. Second, we turn urban street map

Table 3: Parameter and Details

δ	Driving distance between two locations
θ	Driving time between two locations
β	Betweenness score
χ	Closeness score
π	Page Rank score
μ	Average
σ	Std. Deviation
ρ	Correlation



(a) Distance (b) Time

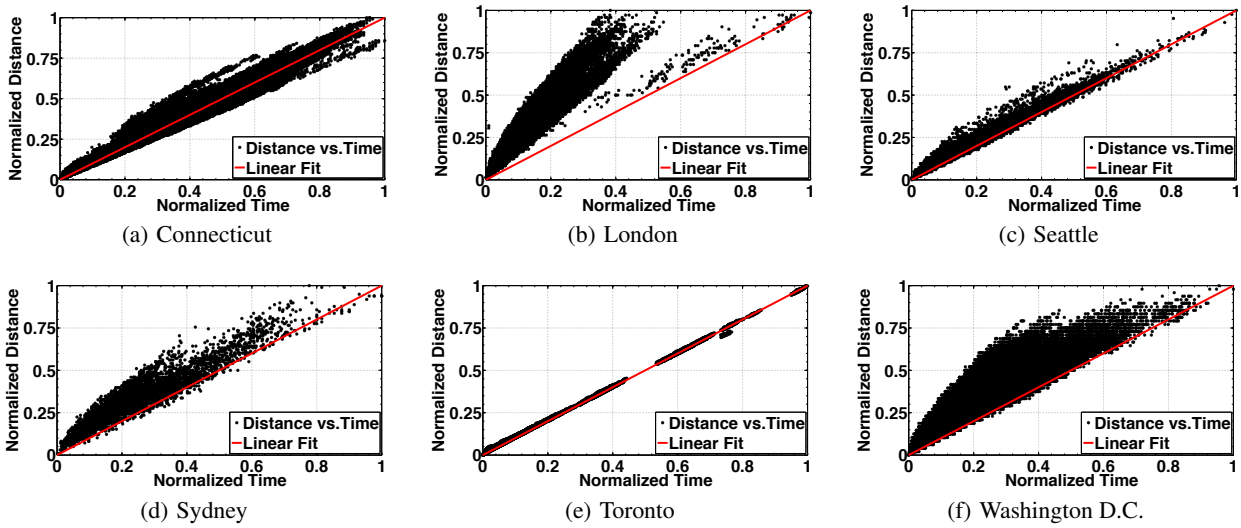
Figure 12: A CCDF of distance and time between locations.

of regions into network graphs, and use measures of centralities to study the structure and function of networks. Here, we want to emphasize the use of travel distance and driving time in location pairs, which helps to focus the analysis only on the motorways, which are taken frequently. This provides an insight into the realistic nature of traffic movements than examining entire cities with infrequent routes [3, 4].

4.1 Distance and Time Analysis

In general the travel distance and time of commuters are significantly influenced by the city size and its interconnection of motorway networks [10]. In case of slow connectivity and congestion, movement shifts to carpooling and rider-sharing approaches.

The first glimpse of the distribution of travel distance and corresponding time in shown in Figure 10. In order to have a perfect correlation between the travel distance and time across all locations of a region, the scatter plot should be centered around the linear fit as visible for Toronto, whose correlation coefficient is 0.97 and shown in Table 4. A good correlation is also found for Connecticut and Seattle where cameras are mostly deployed on highways that have constant speed traffic for long distances. While most of the cameras, which are deployed inside the city of Sydney and Washington D.C. might have more signals and business spots that tend to have slow speed limits and therefore long time to travel short distances, we expect traffic congestion to occur where slow and fast distances meet in these network. We also quantify the cross-correlation between the travel time and distances for the six regions, and their results are shown in Table- 4. The table also provides an insight into the average travel distances and time for these regions.



(a) Connecticut (b) London (c) Seattle

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

Figure 10: A scatter plot for distance vs. time for six regions with linear fit.

Table 4: Results for spatial, temporal and network analysis.

Region	μ_δ	σ_δ	μ_θ	σ_θ	$\rho(\delta,\theta)$	μ_β	σ_β	μ_χ	σ_χ	μ_π	σ_π
Connecticut	60440	34219	42	23	0.88	2259	3679	124	55	0.003	0.0011
London	15605	9737	26	12	0.81	838	1211	98	36	0.005	0.002
Seattle	31126	16427	23	10	0.88	316	477	52	22	0.009	0.004
Sydney	32937	21026	34	16	0.78	127	179	45	16	0.015	0.005
Toronto	125596	148192	75	93	0.97	1879	2659	65	33	0.0048	0.0018
Washington D.C	14611	9040	17	8	0.74	781	1082	186	70	0.004	0.001

Many statistics can be learned from this Table, for example the average deviation in the distance and time. In Figure 12, we show the CCDF of the travel distance and time for the six regions. We find that except Toronto, all other regions have short travel distances, while in Toronto's the average distance is 5 km. In case of travel time, all journeys occur in less than 100 minutes with an exception of Toronto. These results indicate that locations with low correlation are prone to traffic congestion.

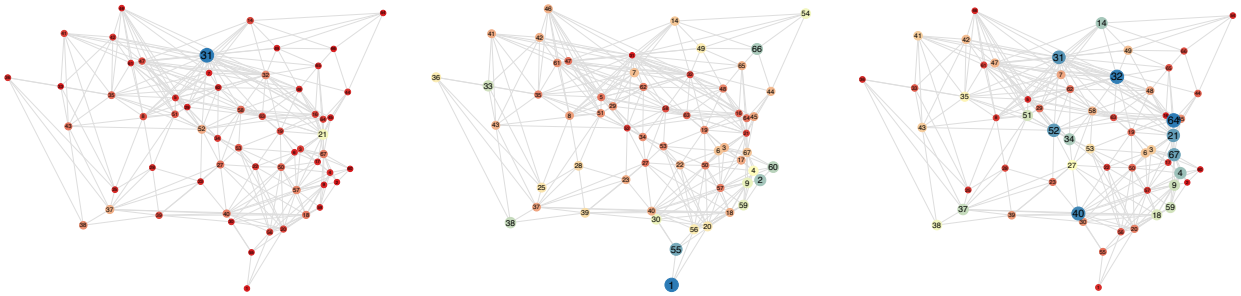
4.2 Network Theory

We examine the structure of motorways using network theory. We represent locations of a region by the vertices of a graph and motorways connecting these locations are represented by the corresponding edges of the graph. In order to study the influence of locations and certain motorways, we count the number of times a motorway is adopted; then assigning the equivalent weight to the corresponding edge in the network graph. Such a representation helps to identify critical junctions and motorways that are prone to congestion, infrequently taken routes and the overall structure of motorway networks.

Network theory has been used in many different places, where the relationships are modeled using network graph and many algorithms are applied on the top of that in order to examine the connectivity, structure, and function of such networks. In this paper, we focus our analysis on measures of centralities that tells, which locations are most central to the network. We are using thee main measures: (i)Betweennness, (ii) Closeness, and (iii) Page Rank. Betweenness centrality measures the extent to which a location lies on a route when moving among other locations. The most visited locations have a high betweenness score and have considerable influence on the connectivity of the road network. Such locations when congested or closed may cause considerable disruption of traffic across the motorway network. In this study, we use the betweenness centrality to discover locations that are highly visited and are deemed important for an efficient route discovery among pairs of source and destination. Closeness centrality measures the mean distance from one location to another. In our case, this centrality helps to identify locations and motorway routes that are infrequently taken and henceforth can aid as alternate routes in case of congestion on the most between locations. These routes are important for evacuation route planning and identifying alternate routes for a city where crisis can bring the traffic to a halt on major routes. We use Page Rank centrality in order to examine the nearest locations that contribute traffic as well as experience the traffic load from nearby locations and motorways. Since formation of congestion is an emergent process, Page Rank centrality can help to identify the tipping points in the network that have the potential to disrupt the traffic at the most between locations and motorways. For more information, interested readers can refer [12].

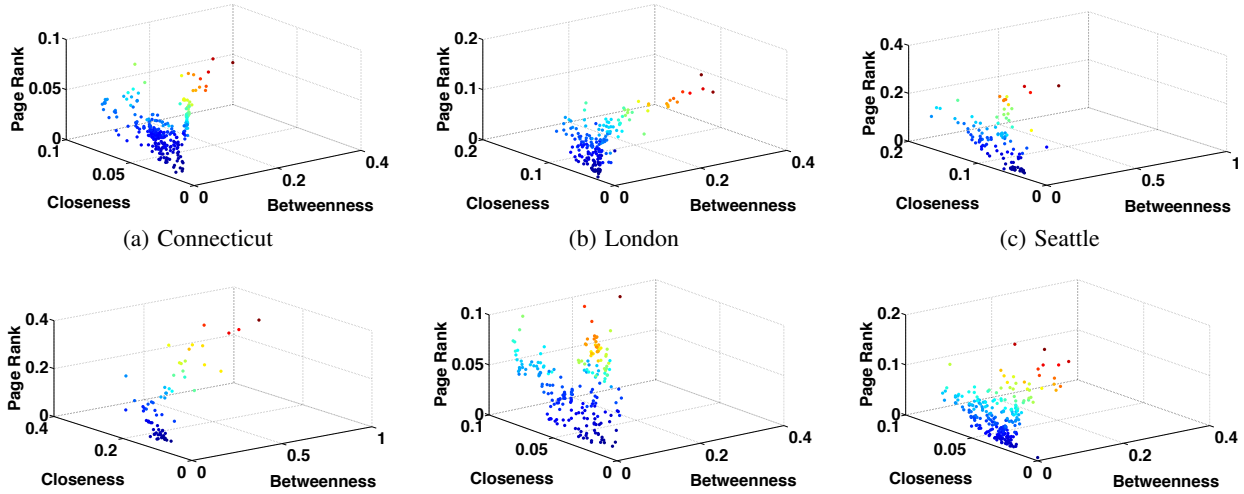
4.3 Network Theory Analysis

We start our analysis by looking at the distribution of various centralities for the city of Sydney in Figure 11. We find that locations 31 and 21 are the most visited locations and their betweenness score is higher than the other locations. It turns out that the location 21 is Sydney bridge that connects two different islands and 31 serves as a major highway (M2) that provides entry and exit points inside the city. In Figure 11(b), the locations that are less frequently visited are the end points of the graphs. Finally, using the page rank centrality, we are able to identify locations that can contribute traffic to most-between locations. As evident, location 64 that is directly



(a) Betweenness (b) Closeness (c) Page Rank

Figure 11: Centrality distribution for Sydney



(a) Connecticut (b) London (c) Seattle

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

Figure 13: The scatter plots show the distribution of three centralities for six regions.

connected to 21 and provides entry and exit to the south eastern part of Sydney. We find similar results for other regions as well. In Table 4, we provide the quantitative numbers for the centralities. In general, we find large deviation from the average values indicating that the networks have skewed connectivity and traffic distribution. We show the distribution of locations with all three centralities in Figure 11. Our results show that in all cities there are at least 2-3 locations that have high very high centralities and henceforth are critical in maintaining the connectivity of the network.

This analysis provides lot of insight into the connectivity of motorways and locations. The results indicate that Sydney, London regions have high betweenness scores and are prone to congestion. On the same lines results of Page Rank centrality show that traffic is emergent in nature and the evidence of traffic present at high betweenness locations can be attributed to the traffic that has passed through the location with high Page Rank centrality. We believe our analysis will open new ways to study the traffic patterns for futures cities and aid in the deployment of ubiquitous systems.

5. SOCIAL ANALYSIS

In this section, we study correlations between the density distribution of pedestrians (humans) through Online Social Networks (OSN) and vehicular traffic. It can be argued that vehicular traffic is a function of human activity in many places such as business centers, museums, and downtowns. Also, human crowd aggregation can aid in developing better prediction models for vehicular traffic congestion. In order to get the data of human activity, we use FourSquare OSN. FourSquare is a location-based social networking website for mobile device users. FourSquare provides users a facility to perform check-ins (mark spatio-temporal presence) at venues (locations they visit, such as restaurants, museums, etc.) in order to help keep up with friends and discover nearby places. We count anonymous check-ins that have been occurred at venues, which are near to the deployed camera locations. This activity helps to quantify the number of humans present in vicinity of cameras.

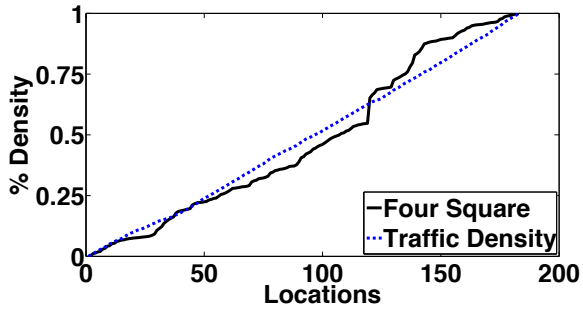


Figure 14: A CDF plot show the distribution of vehicular traffic density to FourSquare check-ins.

We study the distribution of check-ins and vehicular traffic den-

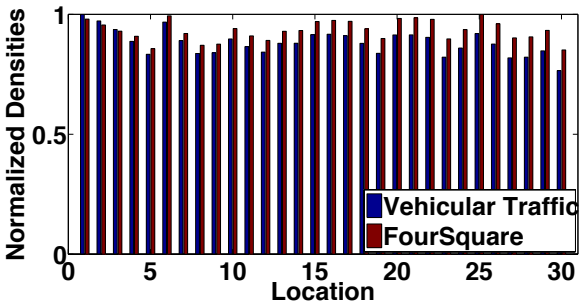


Figure 15: A bar plot that compares the traffic densities to FourSquare check-ins for the city of London.

Table 5: FourSquare Data

Region	Number of Venues	# Checkins
London	9055	1765181
Sydney	3350	20274
Washington D.C	11496	1982339

sity for only those locations, where human activity can takes place. After filtering, we found that three regions, London, Sydney, and Washington D.C. are better suited for this analysis, as most of the cameras are situated in city limits and near business locations. In Table 5, we give the total number of venues and check-ins recorded for these regions. For a fair comparison, the time period of check-ins matches the time period of recorded vehicular densities.

5.1 Analysis

We show the CDF distribution of vehicular density and corresponding FourSquare check-ins at camera locations of London region in Figure 14 (Similar results were observed in Sydney and Washington D.C.). Our results indicate that traffic density and FourSquare check-ins are positively correlated for all the three regions. Although, in case of London, locations around 110 and 150 show some deviation in correlation values and in Sydney some deviation occurs at locations 18 and 60, in general the aggregate results are representative of the assumption that FourSquare checkins are corelated to the traffic densities in the selected urban areas of all three regions. The histogram in Figure 15 gives a distribution of traffic densities against the Foursquare checkins. We find that the results show that human activity is highly correlated (80% for London) with vehicular traffic.

6. CONCLUSION

In this paper, we study the urban infrastructure to enable city-wide ubiquitous computing. We have used the power of global traffic web-cameras, urban street maps and human dynamics to quantify the urban settings of six metropolitan regions around the world. Our vehicular data set has more than 25 million traffic density records, urban street data set has more than 200 thousand routes and human dynamics data is comprised of more than 2 million spatio-temporal check-ins. In this regard, our findings are (i) Urban traffic shows a multitude of traffic patterns beyond the normal rush hour concept. We found regions that initially have no traffic but end up with heavy traffic and vice versa. We also find that vehicular traffic is relatively stable and predictable during weekdays. Historical cities like London show a large deviation in travel distances and time indicating uneven distribution of traffic speed and relatively higher number of signals and shorter routes with several connections. (ii) The network analysis of urban streets indicates that the centrality measures are able to detect frequently visited locations and routes that are prone to traffic congestion. We are also able to detect locations that contribute to emergent traffic congestion. (iii) We find a high correlation between spatio-temporal activity of humans and corresponding vehicular traffic in urban regions such as London and Sydney. We believe our studies will provide a significant insight into the data-driven study of urban infrastructure for enabling city-wide ubiquitous computing. It helps to realize future ubiquitous systems for example that enable vehicle to vehicle and vehicle to road-side type of seamless communication and in identifying where are the communication bottlenecks for a better deployment of computing system. In future, we want to expand our studies to more regions. We also look forward to develop a simulator and provide engineering approaches to ubiquitous computing system based on our experience.

7. REFERENCES

[1] Genevieve Bell and Paul Dourish. Yesterday's tomorrows: notes on ubiquitous computing dominant vision. *PUC*, 2007.

[2] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In *ICPR*, pages 1 –4, dec. 2008.

[3] Alessio Cardillo, Salvatore Scellato, Vito Latora, and Sergio Porta. Structural properties of planar graphs of urban street patterns. *PRE*, 73, 2006.

[4] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality measures in spatial networks of urban streets. *PRE*, 2006.

[5] A Fatah gen Schieck, V Kostakos, A Penn, E O'Neill, T Kindberg, D Stanton Fraser, and T Jones. Design tools for pervasive computing in urban environments, 2006.

[6] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. PH, 2002.

[7] M. Foth, L. Forlano, C. Satchell, and M. Gibbs. *From Social Butterfly to Engaged Citizen*. MIT Press, 2011.

[8] Marcus Foth. Urban informatics, ubiquitous computing and social media for healthy cities. In *MCLC*, 2011.

[9] Yuan-Yih Hsu and Chien-Chuen Yang. Design of artificial neural networks for short-term load forecasting. *GTD*, 1991.

[10] Oded Izraeli and Thomas R. McCarthy. Variations in travel distance, travel time and model choice among smsas. *Journal of Transport Economics and Policy*, 1985.

[11] Sang-Ho Lee, Tan Yigitcanlar, Jung-Hoon Han, and Youn-Taik Leem. Ubiquitous urban infrastructure: Infrastructure planning and development in Korea. *IMPP*, 2008.

[12] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[13] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *PAMI*, 2005.

[14] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1999.

[15] Gautam Thakur, Pan Hui, and Ahmed Helmy. Urban Infrastructure to Enable City-wide Ubiquitous Computing. *Arxiv*, 2013.

[16] Daqing Zhang, Nan Li, Zhi Zhou, Chao Chen, Lin Sun, and Shijian Li. iBAT: detecting anomalous taxi trajectories from GPS traces. In *Ubicomp*, 2011.

[17] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. In *UbiComp*, pages 89–98, 2011.