Analyzing Public Transportation Mobility Data for Networking Purposes

Kais Elmurtadi Suleiman, Otman Basir
Electrical and Computer Engineering
University of Waterloo, Waterloo, Canada
{kelmurta,obasir}@uwaterloo.ca

***Abstract*—Utilizing vehicles for networking purposes has always been a challenge. This is mainly due to the minimum density of connected-vehicles required. The locations of these vehicles should be shareable and reasonably predictable for efficient position-based routing protocols to be implemented. Their Vehicle-to-Vehicle (V2V) communication cooperation should be well-incentivized for efficient networking to be realized. Regular vehicles struggle to have all of these properties. Public transportation vehicles, on the other hand, are well-positioned in this regard; their number is proportional to the number of city residents while being uniformly distributed throughout the day, their locations have no privacy concerns while being highly predictable and their V2V communication cooperation is easily enforceable by the single administration authority they usually have. With efficient networking, public transportation vehicles can become the reliable communication backbone for other vehicle categories. In order to investigate their networking potential, we present for the firs time, in this paper, a data analysis study of realistic public transportation mobility datasets representing the Grand River Transit bus service offered throughout the Region of Waterloo, Ontario, Canada. We show both the data preprocessing and processing phases. The processing phase is mainly based on discovering bus groups using hierarchical clustering. This is done while varying the minimum degree of intra-cluster connectivity and the maximum intra-cluster communication range. Based on this data analysis approach, we show the promising networking potential of public transportation vehicles and provide design guidelines for future networking solutions utilizing them.**

***Keywords*—Public transportation; Mobility; Data analysis; Hierarchical clustering; Vehicular Networking.**

I. INTRODUCTION

Efficient V2V networking can only be realized after a minimum number of connected-vehicles is located within a certain area. The locations of these vehicles should be shareable and fairly predictable to allow for the successful implementation of position-based routing protocols. Mutual cooperation between these vehicles, in terms of sharing their routing resources, should also be well-incentivized. Regular vehicles might struggle to satisfy all of these requirements. On the other hand, public transportation vehicles have the potential to address all of these challenges. Their number is proportional to the number of city residents. Their location distribution is reasonably uniform throughout the day. Their locations have no privacy concerns making them easily accessible while being highly predictable given the nature of public transportation. Their V2V cooperation can easily be enforced by the single administration authority they usually have (e.g. governments). With efficient V2V communication, public transportation vehicles can become the ultimate communication backbone for other vehicle categories.

Given these encouraging properties of public transportation vehicles, we investigate in this paper the networking potential of such means using realistic data analysis. To the best of our knowledge, no previous research work has ever addressed this research gap. The datasets used represent the Grand River Transit bus service offered throughout the Region of Waterloo, Ontario, Canada [1]. Using MATLAB, our data analysis study is divided into two phases: data preprocessing and data processing. In the data preprocessing phase, we first show how the data features have been collected from different source files. Then, we show how both extreme bus waiting times and erroneous bus speeds have been detected and replaced. After that, we show how we have smoothed bus speed data using moving average windows. We also show how we have synthesized the data using linear interpolations between given measurements. This is done in order to guarantee the regular and highly granular data sampling intervals needed in our study. And finally, we show how we have converted location measurements from latitudes and longitudes to Cartesian coordinates in order to use the much faster built-in MATLAB clustering distance function during data processing.

In the data processing phase, we present our approach in terms of evaluating the networking potential using clustering. This approach allows us to measure the number of bus clusters, their cluster size distributions, their average silhouette values and their bus contact durations. These measurements are crucial to our evaluations by corresponding to several ad-hoc networking aspects. As the technique of choice, we use hierarchical clustering while varying the minimum degree of intra-cluster connectivity and the maximum intra-cluster communication range. The minimum degree of intra-cluster connectivity indicates the minimum number of links each bus has with other same-cluster buses. The maximum intra-cluster communication range sets the maximum V2V communication range for buses within the same cluster. Hierarchical clustering allows the variation of these

978-1-5090-6501-1/17/$31.00 © 2017 IEEE

parameters between extremes by varying the linkage method from single to complete and the cutoff distance from 300 meters to a maximum of 1000 meters as set by the DSRC standard [2]. After processing the data, we highlight the promising networking potential identified and provide a set of insights which act as design guidelines for future networking solutions utilizing public transportation vehicles.

The rest of the paper is organized as follows: in Section 2, we overview some related work in order to highlight the novelty of our work and discuss the contributions made afterwards in Section 3. In Sections 4 and 5, we go through the data preprocessing and processing phases, respectively. Based on our data analysis results we outline in Section 6 the design guidelines and tradeoffs for future networking solutions utilizing public transportation vehicles.

II. Related Work

S. Uppoor et. al. in [3], [4] and [5] present a data analysis study of urban mobility data representing regular vehicles throughout the city of Cologne, Germany. The synthetic dataset used has been prepared for studies involving ad-hoc network protocols evaluation. Authors claim that their dataset captures both microscopic and macroscopic mobility aspects of vehicular movement. They show that other datasets lack such scale and realism which results in overestimating Vehicular Ad-hoc NETwork (VANET) protocols.

Using the same synthetic dataset, S. Uppoor et. al. in [6] present probability laws characterizing vehicle mobility between Radio Access Network (RAN) cells. Several networking aspects are evaluated including cell inter-arrival times, cell-residence times and vehicular contact times.

H. Zhu et. al. in both [7] and [8] present a large scale data analysis study using taxi mobility data throughout the city of Shanghai, China. Their study reveals that inter-contact times between vehicles follow an exponential-like tail distribution.

In [9], K. Zhao et. al. use three large scale taxi datasets from the cities of Rome, San Francisco and Beijing. They propose a quad-tree technique to divide these city areas into regions based on the number of taxi visits. These regions are then associated with one of four functions: residential, work, entertainment or other. Efficient delay-tolerant networking solutions can be developed based on these functional regions.

III. Contributions

None of the aforementioned studies offers a mobility data analysis of public transportation vehicles for networking purposes. In [3], [4] and [5], the main focus is on generating a large scale synthetic dataset with a high degree of realism for regular vehicles. The comparison of VANET protocol performances under this dataset and other less-realistic datasets is only used to demonstrate the importance of such realism in avoiding overoptimistic protocol evaluations.

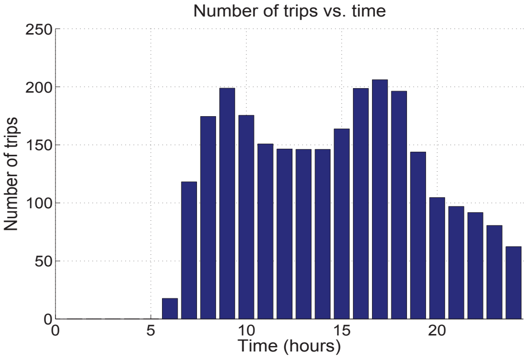


Figure 1: Number of trips vs. time

The focus in [6] is only on characterizing regular vehicle movements within and between RAN cells. The focus of both works presented in [7] and [8] is only on revealing the distribution of inter-contact times between taxi vehicles in urban environments. And finally, the focus in [9] is only about identifying functional regions in order to be utilized by delay-tolerant networking solutions for regular vehicles.

Our work represents the first detailed study of its kind. It first highlights the distinctive and promising V2V networking attributes of public transportation vehicles. It presents a detailed data analysis study using realistic datasets and a novel data processing approach based on hierarchical clustering. It shows the encouraging networking potential of public transportation vehicles and provides a set of design guidelines for future researchers utilizing these vehicles.

IV. Data Preprocessing

A. Data Collection

Data is collected from the sets: bus stop times (9.33 MB file size with 7 features and $252,622$ stop times), bus stops (158 KB file size with 11 features and $2,522$ stops) and bus trips (513 KB file size with 9 features and $6,969$ trips). All of these datasets are available online at [10]. They represent the Grand River Transit bus service offered throughout the Region of Waterloo, Ontario, Canada with an estimated area of $1,046\ km^2$ and a fleet of 259 buses [1]. The stop times dataset offers the features: $trip\ ID$, $arrival\ time$, $departure\ time$ and $stop\ ID$. The stops dataset offers the features: $stop\ ID$, $stop\ latitude$ and $stop\ longitude$. The bus trips dataset offers the features: $service\ ID$ and $trip\ ID$. The $arrival\ time$ and the $departure\ time$ features are converted to numeric values representing fractions of the day. The $service\ ID$ feature is manually encoded as 0 for weekday services, 1 for Saturday services and 2 for Sunday services.

Given these datasets, we compose the overall dataset with the features: $trip\ ID$, $arrival\ time$, $departure\ time$, $stop\ latitude$, $stop\ longitude$ and $stop\ ID$. At this moment, we restrict our study to the much busier trip services offered during weekdays (See Figure 1).

B. Outliers Replacement

We compute the bus $waiting\ time$s by subtracting the $arrival\ time$s from the $departure\ time$s. Then, we draw a boxplot of all bus $waiting\ time$s at each bus stop. Any $waiting\ time$ that is not falling in the range

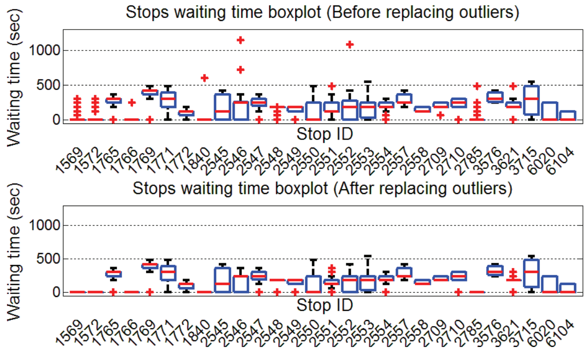


Figure 2: Outliers replacement

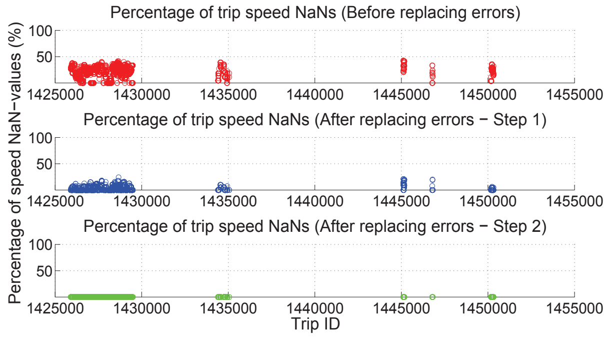


Figure 3: Errors replacement

$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ is considered an outlier where Q_1, Q_3 and IQR are the 1^{st} quartile, the 3^{rd} quartile and the inter-quartile-range of the current bus stop $waiting\ time$ distribution. Such outliers are replaced with the corresponding bus stop $waiting\ time$ distribution medians. Given these new $waiting\ time$s, we recompute the new $departure\ time$s using the same $arrival\ time$s. As shown in Figure 2, we have boxplots of the $waiting\ time$ data both before and after replacing outliers. All stops with initial non-zero $waiting\ time$s are shown.

C. Errors Replacement

We compute the bus $speed$s by first computing the distance between previous and current bus locations and then divide that by the time difference between current bus $arrival\ time$ and previous bus $departure\ time$ given the same $trip\ ID$. We have used the MATLAB function *"lldistkm"* to compute the distance in kilometers between stop locations given their latitudes and longitudes [11]. Buses at first trip stops are assumed to have 0 $speed$s. Any bus $speed$ less than 0 or has a NaN value, is replaced with a moving average of window size 2. After that, we replace the remaining successive abnormal $speed$s with the current trip average $speed$. Given these new $speed$s, we recompute the new $arrival\ time$s and $departure\ time$s using the same distances between previous and current bus locations, the same previous $departure\ time$s and the same current $waiting\ time$s.

Figure 3 shows the percentage of bus $speed$s with NaN values before replacing errors, after the first errors replacement step and after the second errors replacement step. This is shown for all $trip\ ID$s given in the data. Before removing errors, we notice a significant percentage of bus $speed$s with NaN values resulting from having same successive $arrival\ time$s at different locations. These times are measured only in hours and minutes which leads them to look mistakenly the same. After applying the moving average to these abnormal $speed$s (Step 1), we can notice a considerable drop in the percentage of these abnormalities. However, these abnormalities are eliminated completely only after replacing the remaining successive bus NaN-$speed$s with trip average $speed$s (Step 2).

D. Data Smoothing

Given the same $trip\ ID$, we smooth bus $speed$s using a moving average of window size 2. With these new $speed$s, we recompute the new $arrival\ time$s and $departure\ time$s. Figure 4 shows data smoothing applied to the bus $speed$ of a $trip\ ID$ chosen at random. It also shows the $speed$ distribution after data smoothing using the normalized histogram MATLAB function "$histnorm$" [12].

E. Data Synthesis

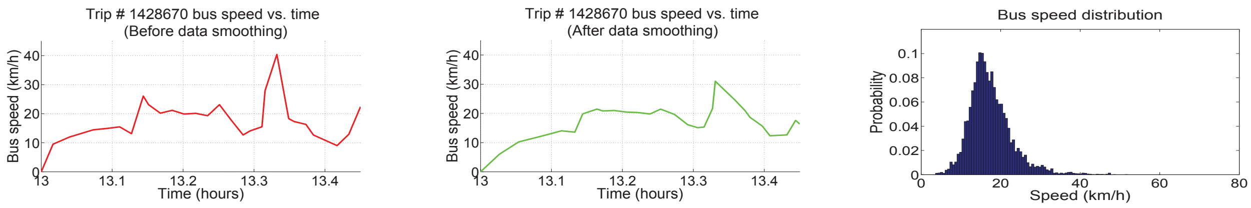
We overcome the irregularity and low-granularity of sample intervals by linearly interpolating between given measurements. The resulting synthetic dataset has measurements generated every 30 seconds for a whole day for each trip. Bus location scatter plots at the least and the most busy hours (i.e. at $6:00$ AM and $5:00$ PM) are both shown in Figure 5. Moreover and in order to verify our data synthesis approach, we show in the same figure the overlap between the original realistic trajectory of a random trip and its trajectory using our synthetic dataset. The MATLAB function **"**$plot\_google\_map$**"** has been used to plot Google maps on figure backgrounds [13].

F. Data Conversion

We convert the resulting synthetic bus $latitude$ and $longitude$ features to Cartesian coordinates in order to use the much faster built-in MATLAB euclidean distance function throughout data processing. Given the area size under consideration, this conversion is made as follows:

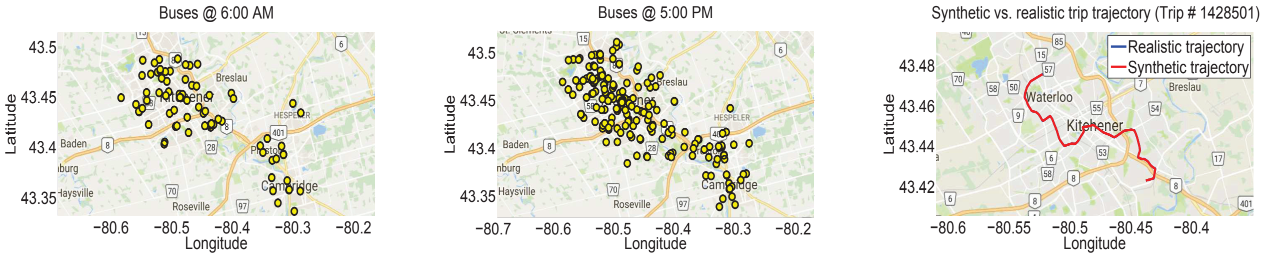
$$b_x = r_{earth} \times cos(b_{lat}) \times cos(b_{lon})$$
$$b_y = r_{earth} \times cos(b_{lat}) \times sin(b_{lon})$$
$$b_z = r_{earth} \times sin(b_{lat})$$

where b_x, b_y & b_z are respectively the bus location x, y & z coordinates, r_{earth} is the earth's radius of 6371 km and b_{lat} & b_{lon} are respectively the bus location latitude and longitude.



(a) (b) (c)

Figure 4: Data smoothing



(a) (b) (c)

Figure 5: Data synthesis

V. Data Processing

A. Approach

To evaluate the networking potential of bus vehicles, we use data clustering to discover bus groups given their b_x, b_y and b_z features. This approach allows us to measure:

- **Number of bus clusters:** which corresponds to the number of cluster heads; having less cluster heads increases the chances of having congestions. However, it would also mean less reliance on other network tiers if these cluster heads were to act as gateways.
- **Bus cluster size distribution:** which shows the hourly bus cluster size distribution using boxplots; having cluster sizes with larger means or more cluster members increases the chances of congestions at cluster heads. Moreover, having cluster sizes with larger inter-quartile ranges means less fairness since some clusters will have more members and therefore a higher chance of experiencing congestions and a longer time for messages to reach all members.
- **Bus cluster average silhouette value:** which shows the degree of location similarity between same-cluster buses; less similarity means a longer distance/time and a higher number of hops for messages to reach all cluster members. This is independent from the bus chosen as the cluster head.

As we shall see in the next subsection, we use hierarchical clustering for the following reasons:

- It allows us to vary the minimum degree of intra-cluster connectivity by adjusting the linkage method from single to complete. Single linkage corresponds to less communication reliability where each bus has at least one connection with another same-cluster bus member. Complete linkage corresponds to more communication reliability where each bus is completely linked to other same-cluster bus members. This complete linkage might be needed under harsh channel conditions, and
- It allows us to vary the maximum intra-cluster communication range from 300 to 1000 meters by adjusting the cutoff distance.

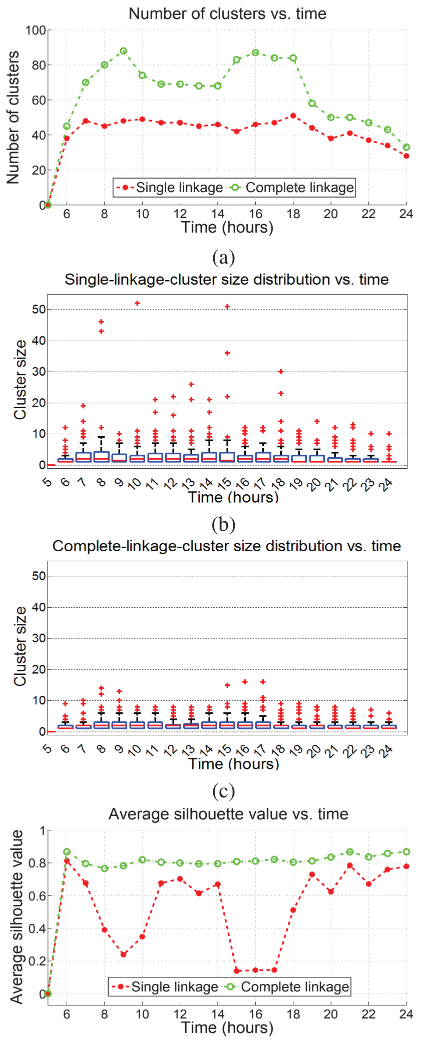
Other well-known reasons to use hierarchical clustering include: its ability to deal with non-globular data, its non-reliance on foreknowing the number of clusters, its consistency and insensitivity to initial clustering parameter assumptions, its robustness to outliers and data density variations, and its acceptable clustering speed given its capability to utilize parallel computing. However, we do not claim that hierarchical clustering is the best/only technique applicable but it suffices the purposes of our study.

We evaluate the networking potential of buses by also computing contact durations. This is done while adhering to a varying maximum intra-cluster communication range (from 300 to 1000 meters) given the features: b_x, b_y and b_z. The goal is to measure:

- **Bus contact duration distribution:** which shows all daily contact durations, their mean value, 1^{st} and 3^{rd} quartiles; longer durations correspond to better communication.

B. Experiments

1) Varying the Minimum Degree of Intra-cluster Connectivity: Assuming the intra-cluster communication range of 1000 meters, we vary the minimum degree of intra-cluster connectivity using hierarchical clustering by adjusting the linkage method from single to complete. Figure 6 shows the effects of this variation. The number of bus clusters varies from 33 to 88 under complete linkage with an average of 62.5 while varying at a lower range,



(a)

(b)

(c)

(d)

Figure 6: Varying the Minimum Degree of Intra-cluster Connectivity

from 28 to 51, under single linkage with an average of 41.05. The bus cluster size distribution under complete linkage is skewed reaching as high as 16 buses/cluster while still exhibiting skewness under single linkage with as high as 52 buses/cluster. The average silhouette value of bus clusters reaches as high as 0.8685 under complete linkage and as low as 0.1389 under single linkage.

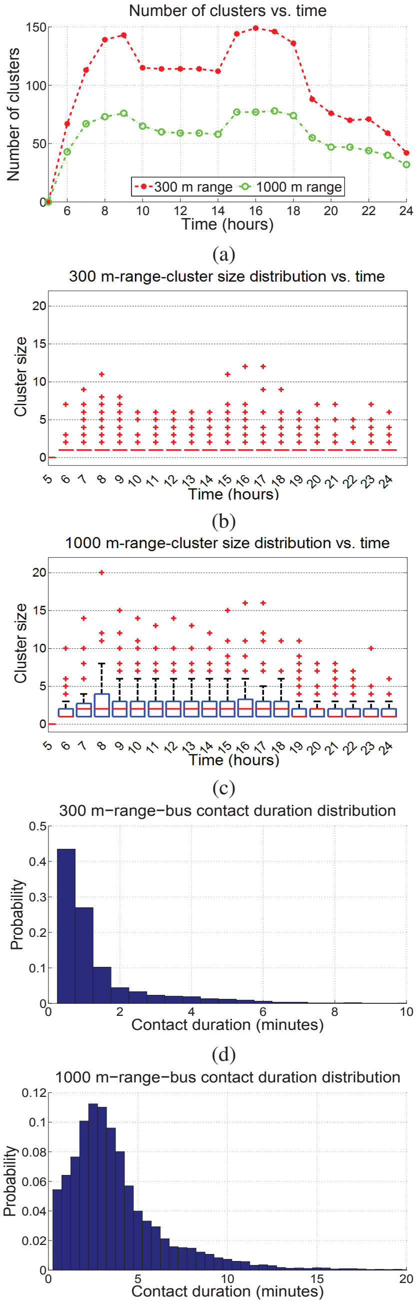
2) Varying the Maximum Intra-cluster Communication Range: Assuming the average linkage method, we vary the maximum intra-cluster communication range using hierarchical clustering by adjusting the cutoff distance from 300 to 1000 meters. Figure 7 shows the effects of this variation. It also shows the bus contact duration distributions under both 300 and 1000 meter bus communication ranges. The number of bus clusters varies from 42 to 149 under 300 meter cutoff distance with an average of 100.6 while varying at a lower range, from 32 to 78, under 1000 meter cutoff distance with an average of 56.55. The bus cluster size distribution under 300 meter cutoff distance is skewed reaching as high as 12 buses/cluster while still exhibiting skewness under 1000 meter cutoff distance with as high as 20 buses/cluster. The bus contact duration has a mean as high as 1.38 minutes, with $Q_1 = 0.5$ minutes and $Q_3 = 1.5$ minutes, under 300 meter cutoff distance and an even higher mean of 3.89 minutes, with $Q_1 = 2$ minutes and $Q_3 = 4.5$ minutes, under 1000 meter cutoff distance.

VI. Discussion

As shown by the results, the networking potential of public transportation vehicles, represented by buses in our study, is promising. Resulting bus clusters are much fewer than the number of bus trips, a significant some of their sizes corresponds to high proportions of these trips, their average silhouette values show closeness under complete linkage and wide spread under single linkage especially at busy hours, and a lot of their bus contact durations are long enough to support reasonable V2V communication. We can also see that deciding the minimum degree of intra-cluster connectivity or the maximum intra-cluster communication range is a matter of designer preference and application requirement. This is since:

- Single linkage has led to: fewer clusters, bigger cluster sizes, larger cluster size inter-quartile ranges and lower average silhouette values. This means fewer cluster heads, less reliance on other network tiers, more cluster head congestions, less fairness and more time for messages to reach all cluster members. This is in addition to less channel redundancy and therefore reliability,
- Complete linkage has led to: more clusters, smaller cluster sizes, smaller cluster size inter-quartile ranges and higher average silhouette values. This means more cluster heads, more reliance on other network tiers, fewer cluster head congestions, more fairness and less time for messages to reach all cluster members. This is in addition to more channel redundancy/reliability,
- 300 meter communication range has led to: more clusters, smaller cluster sizes, smaller cluster size inter-quartile ranges and shorter contact durations. This means more cluster heads, more reliance on other network tiers, fewer cluster head congestions, more fairness and less V2V communication chance,
- 1000 meter communication range has led to: fewer clusters, larger cluster sizes, larger cluster size inter-quartile ranges and longer contact durations. This means fewer cluster heads, less reliance on other network tiers, more cluster head congestions, less fairness and more V2V communication chance.

This set of observations act as guidelines to network designers utilizing public transportation vehicles. It does so by giving insights into the extreme cases of: single linkage, complete linkage, 300 meter communication range and 1000 meter communication range. Based on these insights, designers can choose their varying/fixed linkage method and communication range while meeting their priorities.



(a)

(b)

(c)

(d)

(e)

Figure 7: Varying the Maximum Intra-cluster Communication Range

VII. Conclusions

Public transportation vehicles have the potential to act as a reliable communication backbone for other vehicle categories due to their numbers, uniform city distribution, their shareable and highly predictable locations and their enforceable V2V communication cooperation. This is in contrary to regular vehicles which struggle to have all of these properties. In this paper, we have presented, for the first time, a study of public transportation mobility data for networking purposes using realistic bus mobility datasets. We have presented in details the two phases of data preprocessing and processing. Our novel approach in the data processing phase is mainly based on evaluating the buses networking potential in terms of their number of clusters, cluster size distributions, cluster average silhouette values and bus contact durations. We mainly use hierarchical clustering for this purpose while varying the minimum degree of intra-cluster connectivity, by varying the linkage method from single to complete, and the maximum intra-cluster communication range, by varying the cutoff distance from 300 to 1000 meters. Results have demonstrated the promising networking potential of public transportation vehicles represented by buses in this study. Based on these results, a set of insights have been made which act as guidelines for future network designers utilizing public transportation vehicles.

References

[1] Wikipedia, “Grand River Transit”. Available: https://en.wikipedia.org/wiki/Grand_River_Transit

[2] J. B. Kenney, "Dedicated Short-Range Communications (DSRC) Standards in the United States," in Proceedings of the IEEE, vol. 99, no. 7, pp. 1162-1182, July 2011.

[3] Sandesh Uppoor and Marco Fiore. 2012. Insights on metropolitan-scale vehicular mobility from a networking perspective. In Proceedings of the 4th ACM international workshop on Hot topics in planet-scale measurement (HotPlanet ’12). ACM, New York, NY, USA, 39-44.

[4] S. Uppoor and M. Fiore, "Large-scale urban vehicular mobility for networking research," 2011 IEEE Vehicular Networking Conference (VNC), Amsterdam, 2011, pp. 62-69.

[5] S. Uppoor, O. Trullols-Cruces, M. Fiore and J. M. Barcelo-Ordinas, "Generation and Analysis of a Large-Scale Urban Vehicular Mobility Dataset," in IEEE Transactions on Mobile Computing, vol. 13, no. 5, pp. 1061-1075, May 2014.

[6] S. Uppoor and M. Fiore, "Characterizing Pervasive Vehicular Access to the Cellular RAN Infrastructure: An Urban Case Study," in IEEE Transactions on Vehicular Technology, vol. 64, no. 6, pp. 2603-2614, June 2015.

[7] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu and L. M. Ni, "Impact of Traffic Influxes: Revealing Exponential Intercontact Time in Urban VANETs," in IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 8, pp. 1258-1266, Aug. 2011.

[8] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li and L. M. Ni, "Recognizing Exponential Inter-Contact Time in VANETs," 2010 Proceedings IEEE INFOCOM, San Diego, CA, 2010.

[9] K. Zhao, M. P. Chinnasamy and S. Tarkoma, "Automatic City Region Analysis for Urban Routing," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 1136-1142.

[10] Transit - GRT GTFS Static & GTFS-realtime Data Feed. Available: http://www.regionofwaterloo.ca/en/regionalGovernment/GRT_GTFSdata.asp

[11] M. Sohrabinia, “Find distance between two points based on latlon coordinates”. Available: https://www.mathworks.com/matlabcentral/fileexchange/38812-latlon-distance

[12] A. Serrano, “Same as histogram, but the area sum is 1”. Available: https://www.mathworks.com/matlabcentral/fileexchange/22802-normalized-histogram

[13] Z. Bar-Yehuda, “Plot a google map on the background of the current figure using the Static Google Maps API”. Available: https://www.mathworks.com/matlabcentral/fileexchange/27627-zoharby-plot-google-map