Carpooling Recommendation System

Adeela Huma
Dept. Of Computer Science
Virginia Tech
Falls Church, VA 22043
ahuma@vt.edu

Ahmed Elbery
Dept. Of Computer Science
Virginia Tech
Blacksburg, VA 24060
aelbery@vt.edu

Antonio Fuentes
Dept. Of Civil & Env. Engineering
Virginia Tech
Blacksburg, VA 24060
fantonio@vt.edu

ABSTRACT

Carpooling is known to have a broad range of benefits to individuals, organizations and governments. However, carpooling is not efficiently applied because of the lack for manageability and trustworthiness. In this paper, in order to minimize the impact personal vehicle usage has on commuters, city transportation networks and the environment; a carpooling recommendation system for drivers is presented. The approach and methodology to the development and implementation of this carpooling recommendation system is also described. The dataset utilized is referred to as Gowalla, it is a location based online social network obtained through Stanford University. The data provided consists of spatio-temporal information as well as a friendship network of participating users. A general framework consisting of a database including the previously mentioned data, a scheduling subsystem and a recommendation subsystem in proposed. The primary focus discussed in this paper consists in addressing the recommendation subsystem. The recommendation subsystem can be broken down into three stages, which are similarity detection, history/prediction and recommendation. To find user similarities, the relation between location and users as well as the social network between the users was applied. Furthermore, to model/predict the mobility continuous time Markov Chain (CTMC) is executed. The result shows that the number of carpool matching depends on many parameters such as the visiting probability threshold. The result shows that the matching ratio ranges between 96% and 34% when the parameter α is between 1 and 2.

Keywords

Carpooling recommendation, geo-location check-ins, friendship networks, Markov Chain, k-nearest neighbors

1. INTRODUCTION

An important and chronic problem facing us now is the global warming coupled with pollutant emissions. In 2008, the U.S. Department of Energy mentioned in [1] that approximately 30% of the fuel consumption in the U.S. is consumed by vehicles moving on the roadways. In addition, about one-third of the U.S. carbon dioxide (CO_2) emissions come from vehicles [2]. The 2011 McKinsey Global Institute report estimated savings of "about $600 billion annually by 2020" in terms of fuel and time saved by helping vehicles avoid congestion and reduce idling at red lights or left turns. In this context, carpooling can play an important role in mitigating these problems by minimizing the number individual trips.

Carpooling covers many objectives that span different levels, such as: individual, organizational and governmental, and international levels. From an individual point of view, carpooling can save money and the effort of driving. This means that individuals have self-motivations to carpool with others. Also carpooling saves fuel and reduces road congestion, which the national government has as objectives to address. For organizations, carpooling helps overcoming the shortage of parking slots, which different organizations spend large amounts of money to address. Reducing the pollutant emissions also help solving global warming issues.

Because of these reasons, there are many efforts to encourage travelers to carpool. Some of them are imitated by individuals who established social media groups to find carpool matching among the group users. This is a good direction, however it lacks for scalability, trustworthiness and manageability. The scalability is limited to the group users only and the locality of those users. In addition, carpooling requires a minimum level of mutual trustworthiness among those potential carpoolers. However, the social ties on social media may be unaffected under the trustworthiness constraint because most of the social media users have privacy concerns. A third obstacles facing this direction is the manageability problem, where users need to search for the carpooling matches, then agree about the location and schedules, as well as negotiating the costs.

Another direction is the centralized direction, which takes a good phase from social media groups. For example, the "Go! Vermont" program [3], which was launched by the state of Vermont, is a centralized online system for finding carpool matching. On this system the users should register and schedule their trips or search for carpool matching. The system also provides a tool for cost estimation. The main drawback here is that the system is reactive, which means the users have to schedule or search a carpool matching.

At the governmental level, the High Occupancy Vehicles (HOV) lanes [4-6] have been established by the USA government on some roads and dedicated to vehicles with 2+ commuters. In this way, the government uses this technique as an incentive to encourage the commuter to carpool. However, some studies on the effectiveness of these HOV lanes such as [5] showed that these HOV lanes have negative impact on the congestion in the roads because it reduces the road capacity. The reason is that these HOV lanes are underutilized most of the times while the lanes adjacent to them are heavily congested.

Another important issue is that neither of these systems or efforts utilize on new technologies such as vehicular communication and big data. For instance, social online networks are rich sources of data that can be analyzed and mined to get valuable information. This immense amount of information about users, their interests and relationships could be used in many applications. As wireless communications advance, new technologies are being imbedded such as GPS, such technologies provide location information about users. There are datasets that describe the user trajectories such as Microsoft data set [7] which records the trajectories of users using multiple transportation methods (Walk, Taxi, Car…..). However, due to privacy consideration, there is a lack of data that integrates user information and user mobility and location information.

2. OBJECTIVE

Despite all these efforts made to encourage commuters to carpool, commuters find it difficult to locate people willing to share trips, and to make the arrangements because most carpooling efforts are made ad-hoc using social media groups, on a small level.

The objective of this project is to produce an automated, easy to use system for efficient carpooling. The proposed system utilizes on new technologies such as VANET communication, social networks, big data and data analysis techniques to find the best carpool matching.

A main difference between the proposed system and current systems is that, it should be smart enough to find user similarities and predict future user trips based on the mobility history of the users. Based on those prediction and similarities, the system can find the matches for carpooling.

Therefore, the objective is to build an automated proactive carpooling system that will maximize the number of carpooling and consequently will minimize the number of individual trips. And thus it would reduce fuel consumption, environmental impact, as well as reduce the traffic congestion and travel times.

3. SYSTEM GENERAL FRAMEWORK

The general framework is depicted in Figure 1. It incorporates user trips schedule through the scheduling subsystem with the user mobility history stored in the mobility database. It also uses the social network for users, which is stored in the friendship database.

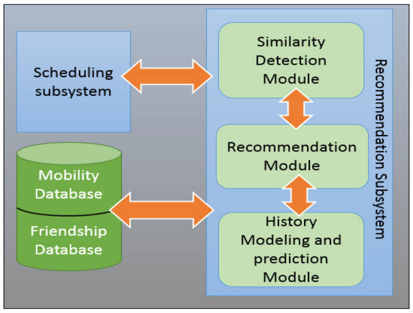


Figure 1. General Framework for the Carpooling Recommendation System

Based on these data the system applies learning, modeling and prediction techniques to: -

a) Find similar users
b) Predict the future user's trips
c) Create carpooling recommendation for matched users

To generate the recommendations, the data can flow in either of two directions as shown in Figure 2.

The recommendations can be initiated by a unique user who schedules a trip to a specific location within a time window. From which the system finds the similar users and friends to the unique user and sends them to the recommendation module, which asks the prediction module. Which in its turn estimates whether those user willing to visit this venue within this time interval, and return this estimation to the recommendation module to generate the recommendations.

The other flow starts periodically by the system itself. The system periodically predicts the future user visits and sends this information to the recommendation module, which finds similar users and recommends them to carpool together.

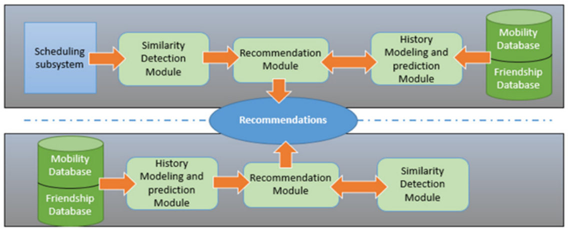


Figure 2. Information Flow in the System

4. DATA SELECTION AND ANALYSIS

The dataset utilized is referred to as Gowalla, and it is a location based online social network obtained through Stanford University's, Stanford Network Analysis Project (SNAP) [8], under the Stanford Large Network Dataset Collection. The data provided consists of a geo-location bipartite graph and a friendship network edge list. The geo-location bipartite graph consisted of the following information listed below, while the friendship network consisted of an edge list that represented the friendship between the users in the aforementioned bipartite graph. Overall there are 196,591 nodes, 950,327 edges and 6,442,890 check-ins in the dataset [8]. The check-ins has these fields User_id, Check-In Time, Latitude, Longitude and Location Id

Utilizing the software tool and power of ArcGIS, the check-in bipartite graph was able to be plotted in a spatial map through the use of latitude and longitude coordinates that were provided. The majority of the check-ins where determined to be in the United States and Europe as can be seen in Figure 3.

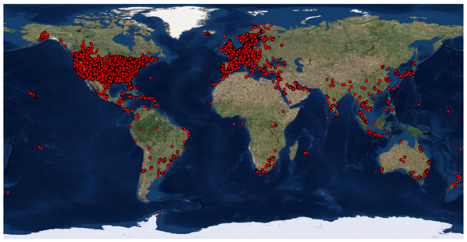
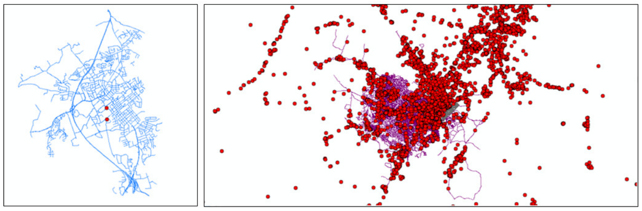


Figure 3. Complete Gowalla check-ins data

In order to have a better understanding as well a better ability to assess the proposed carpooling recommendation system. A subset of data was composed based on an "area of interest" determined by the authors. Two locations where initially considered, the first was Blacksburg, Virginia where the main campus of Virginia Tech is located and the second was the Northern Capital Region of Virginia Tech located in Northern Virginia. In relation to the available data, that can be seen in Figure 4 it was determined that the Northern Virginia area which includes Washington D.C and Maryland would be used.

Thus the "area of interest" new subset consisted of 19,840 rows of data of which 479 where unique users and 6,700 unique location IDs, Figure 5 illustrates the "area of interest". Although the number of data rows is still quite large, many of these check-ins were found to have behavior that might influence the carpooling system in a negative way. Such as check-ins by the same user in the same location with minimal time intervals. For example, a case where a user checks-in at location A at 5:00 PM and does so again at 5:05 PM.



Blacksburg, Virginia

Northern Virginia

Figure 4. Check-in comparison between Blacksburg, Virginia and Northern Virginia

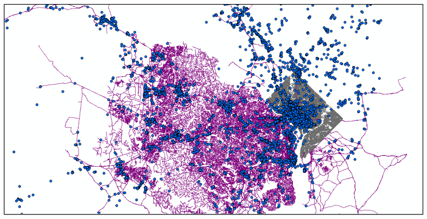


Figure 5. Final area of interest: Northern Virginia, Washington D.C, Maryland

Furthermore, a visual spatial analysis was conducted to visualize and assess how the check-in network behaved in terms of one specific location. Thus, the unique location ID corresponding to the white house located in Washington D.C. was utilized as the specific location. It was determined that the white house had a total of 64 unique users check in. And through the friendship edge list provided, there were a total of 4,261 unique friendships to those 64 users.

The next procedure consisted of expanding the reduced network to investigate the connection between the users that checked-in at the white house and the users listed in the friendship network. It was determined that this subset of data had 170 unique users and 2,874 unique location ID's. This subset increased from the previous subset because it now includes additional check-ins made not only by those users which checked-in at the white house, but also the friends of those users.

Refining the dataset further, of the users who checked into the white house, 29 of them were found to be connected through the friendship network. Therefore, this is an important distinction to make because this data can be used to identify the overall tracks of similar users and thus, possibly provide a carpooling recommendation between similar locations. In the figure below the yellow nodes represent the 29 user check-ins from the white house connected through the friendship network and their spread through the "area of interest".

5. SYSTEM MODEL AND DEVELOPMENT

This section describes the details about the system model. First it shows the problems in the dataset and how we did overcome them. Then the similarity detection module and its implementation is presented. And finally the mathematical foundation and the implementation for the mobility modeling and prediction using continuous time Markov Chain CTMC is presented.

5.1 Data Preprocessing

The first step involved in pre-processing data was to identify an urban area where we can test our car pool recommendation approach. The area of Virginia, DC, Maryland was chosen as illustrated in Figure 6.

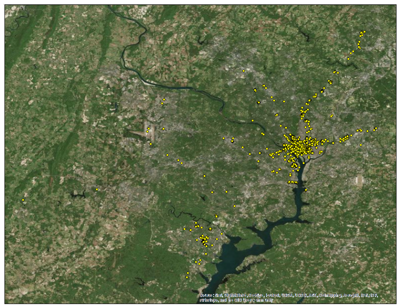


Figure 6. Users connected through the friendship network that checked in at the white house within "area of interest".

Some issues were encountered with the dataset, one was multiple check-ins within short intervals of time at the same location. This could be problematic in prediction when considering the temporal aspect of data. A minimum check-in interval of an hour was defined. Any check-ins within that period at same location was considered one check-in. Another issue in the dataset was that same location had multiple location Ids. This could give the wrong impression by modeling a user that has been to multiple locations when in fact it's within a certain range of same location. The area of interest was divided into small regions each of square 500 m X 500 m. All location ids within one square were considered as one location. An additional issue was users associated with few check-ins (< 2). Such users would not show up in search due to lack of data. Therefore it was decided a threshold value to track active users only and remove inactive users.

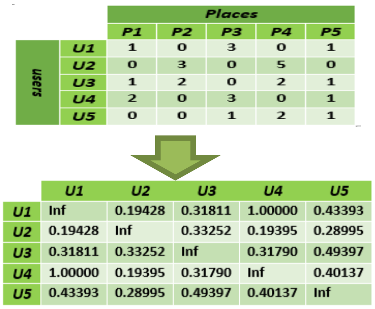
5.2 User Similarity Detection

The goal of similarity detection module is to find which users are the nearest match for a given user. By nearest match it is intended that if user X is traveling from location A to location B, it would be determined which other users have high probability of also going from and to location A and B respectively.

In the Gowalla dataset, check-in location is considered as destination location for user. Since carpool recommendation needs to be done between some source and destination locations. Source location for a user seems to be missing in the Gowalla cheek-in dataset. Ideally source location will represent the home address of a user, and similarity detection is not possible without that. Home addresses for users was unavailable, thus using the Gowalla friendship network clusters were created of users using k-mean clustering and assign each cluster a location Id that was assigned to all users in the same clusters as their source location Id. We are assuming that friends who are in one cluster are considered in the same region hence the same source location Id.

After having the key features for similarity detection (user Id, source location Id, destination location Id). A bipartite graph was created between users and source location Id. Edge weight represent the frequency of visit to place 'j' by a user 'i'. From this bipartite graph we calculated the distance matrix between users using k-nearest neighbor approach. This distance matrix gives information about which user(s) is similar to other user(s).

Another approach that can also help in similarity detection based on location creating a distance matrix between places using the above bipartite graph as shown in Figure 8. This measure will give information that which places are closer.



		Places				
		P1	P2	P3	P4	P5
users	U1	1	0	3	0	1
	U2	0	3	0	5	0
	U3	1	2	0	2	1
	U4	2	0	3	0	1
	U5	0	0	1	2	1

	U1	U2	U3	U4	U5
U1	Inf	0.19428	0.31811	1.00000	0.43393
U2	0.19428	Inf	0.33252	0.19395	0.28995
U3	0.31811	0.33252	Inf	0.31790	0.49397
U4	1.00000	0.19395	0.31790	Inf	0.40137
U5	0.43393	0.28995	0.49397	0.40137	Inf

Figure 8: User Check-ins and Similarity Matrix

5.3 User Modeling and Mobility Prediction

The mobility modeling and prediction is determined through the application of continuous time Markov Chain (CTMC) [9]. Important characteristics to mention are that the CTMC assumes exponential distribution of transition times and irreducibility of the chain. The CTMC is fully characterized by the transition rate matrix (Q-matrix).

$$Q = \begin{bmatrix} q_{1,1} & q_{12} & \cdots & q_{1,n} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n,1} & q_{n,2} & \cdots & q_{n,n} \end{bmatrix}$$

Where $q_{i,j}$ is the transition rate from state i to state j. And $\{v_i\}_{i\in S}$ is the parameter of the exponential distribution of sojourn time T_i

$$T_i = \frac{1}{v_i} \;, \qquad v_i = \sum_{j \neq i} q_{i,j}$$

Furthermore, the state probability $\pi_i(t)$ is defined as the probability that the system is in state i at time t is

$$\pi_i(t) \;=\; P\{X(t) \;=\; i\}$$

And the probability distribution $\pi(t)$ is the vector of the state probabilities for all the n states.

$$\pi(t) \;=\; \big(\pi_1(t),\, \pi_2(t),\; \ldots\; \pi_n(t)\big)$$

Given the exponential distribution of time, the state probability at any time of irreducible Markova chain depends only on the current state and the Q-matrix as shown in the following equation.

$$\pi(t) \;=\; \pi_0\, e^{tQ}$$

Based on this mathematical foundation, the user mobility can be modeled using continues time Markov chain. The states of the chain represents the different places he visited, and the time between consequent transitions is used to create the transition rate matrix (Q-matrix).

Therefore, to model the user mobility using CTMC it must be made sure that:-

a) T_i is exponentially distributed
b) The chain for user state is irreducible

For the first condition, Figure 9 shows the distribution of the transition time for some users. The figure shows that the distribution of transition time is very close to exponential distribution with small deviations. Thus it is safe to use.

The irreducibility of the Markov chain simply means that the chain is a connected graph (i.e. from any state there is a path to any other state). If the chain is not irreducible this means that the system becomes stuck at a state and cannot move any other state. Also the irreducible chain converses to a state probability distribution as time goes to infinity.

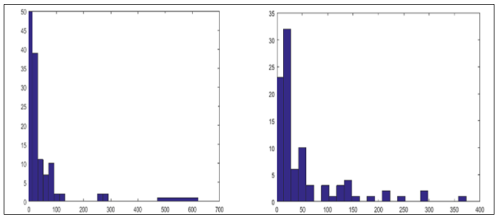
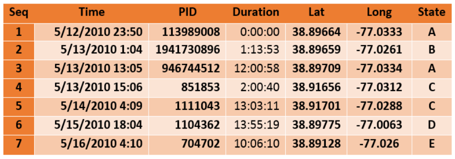


Figure 9. The Distribution of T_i for

Figure 10 shows a user trajectory in tabular format to give the details of his mobility. Then Figure 11 shows his mobility trajectory on the map and his state transition diagram.



Seq	Time	PID	Duration	Lat	Long	State
1	5/12/2010 23:50	113989008	0:00:00	38.89664	-77.0333	A
2	5/13/2010 1:04	1941730896	1:13:53	38.89659	-77.0261	B
3	5/13/2010 13:05	946744512	12:00:58	38.89709	-77.0334	A
4	5/13/2010 15:06	851853	2:00:40	38.91656	-77.0312	C
5	5/14/2010 4:09	1111043	13:03:11	38.91701	-77.0288	C
6	5/15/2010 18:04	1104362	13:55:19	38.89775	-77.0063	D
7	5/16/2010 4:10	704702	10:06:10	38.89128	-77.026	E

Figure 10. Example for User Trajectory

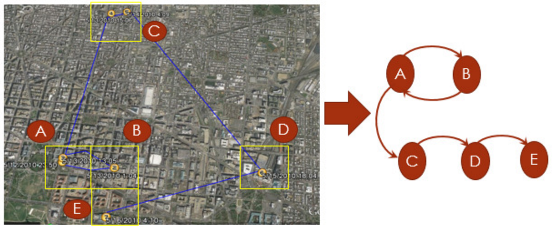


Figure 11. Example for User Trajectory and State Transition Diagram

It is clear that the chain is not irreducible. For example if the user comes to state C, D or E, he would not be able to return to neither states A or B.

This chain state diagram should be converted to an irreducible chain. To do that we use the idea of home state. The main idea is that, if a user visited two different places in two different days, then it is intuitive that he returned back to his home between the two visits. The home here is not necessary to be his home address, it may be a hotel or any place in which the user sleep or takes a rest. So this state is not a real location it is just a representation to rest between two visits. Figure 12 shows the chain after adding the new home state (H). It is clear that the chain become irreducible.

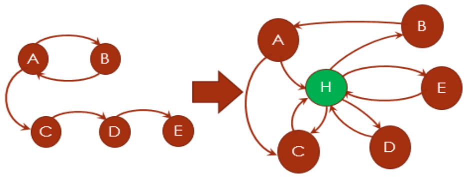
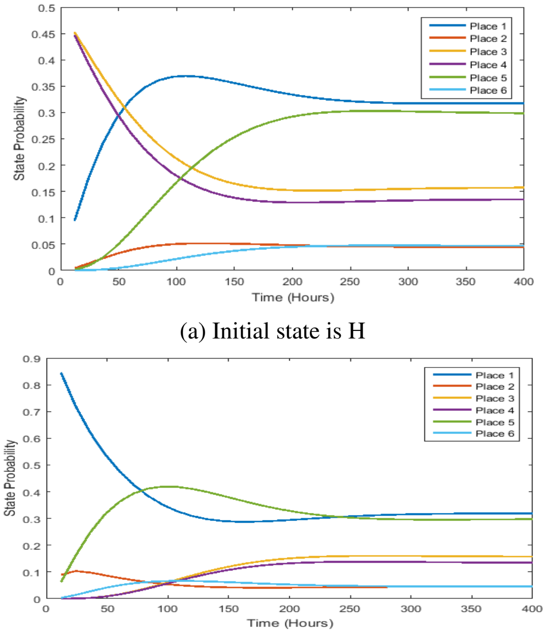


Figure 12. Final Chain State Diagram

6. Experimental Results.

Figure 13 show the state probability distribution for some user and how it changes with time. It also shows the impact of the initial state. The number of matching found for each user depends on some parameters. One main parameter is the visiting probability threshold.



(a) Initial state is H

(b) Initial state is Place #1

Figure 13. State probability distribution for user 88

We define this threshold for each user as multiples of the average visiting probability

$$p_{thresh} = \frac{\alpha}{number\ of\ places\ visited\ by\ the\ user}$$

Where α is an integer. In this way, p_{thresh} represents the user's average visiting probability to any place. Following, Figure 14 shows the relation between number of matching found and the parameter α, it shows that the number of matching decreases exponentially with α. As shown, when α is between 1 and 2 a reasonable value that results around 60% matching is found. The result shows that the matching ratio ranges between 96% and 34% when the parameter α is between 1 and 2.

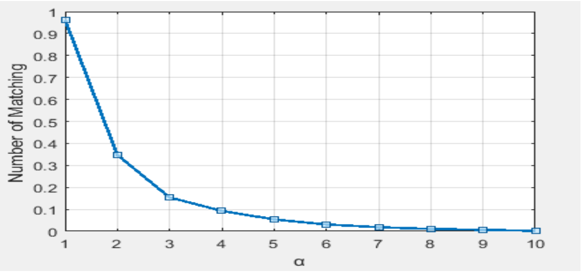


Figure 14. Number of Matching vs α

7. SUMMARY AND CONCLUSION

In conclusion, the necessity of a carpooling recommendation system is a major benefit to many aspects of society ranging from personal benefit, government decision-making and environmental impacts. The idea of a recommendation system that successfully provides carpooling recommendation based on user, location, time and user friendship.

Spatial-temporal information from the Gowalla [8] data set was utilized to determine the interaction between users, locations and the friendship between users. From the original data set, a subset of data was extracted based on spatial interests and labeled the "area of interest." Small analysis of comparison was conducted based on user locations and the spread throughout the "area of interest".

Utilizing the aforementioned subset of data, k-nearest neighbor approach was used to detect similarity between users and the continuous time Markov Chain (CTMC) was used for user modeling and mobility prediction. Overall the goal of recommending carpooling between users by taking into account their location was determined to range between 96% and 34% when the parameter α is between 1 and 2.

8. Author Contributions

Each author contributed a significant amount of work based on his or her skills and contributed to the final product. Adeela Huma addressed the data processing and user similarity analysis. Ahmed Elberry addressed the problem statement, literature background and fundamental analysis of user modeling and mobility prediction. Antonio Fuentes plotted, refined and provided the original data and subsets of data for use throughout the analysis as well as provided GIS analysis and relationships. Each member contributed equal and beneficial effort to the final presentation and final report.

9. Code URLs

All the code used and corresponding data files can be found at :

https://drive.google.com/folderview?id=0B01yo9YbQia1ZHg1RHI2X0dpRFU&usp=sharin

To run it, just run the Run.m file in matlab.

10. REFERENCES

[1] U. S. D. Energy, "Annual Energy Outlook 2008, With Projection to 2030,," *Energy Inf. Admin., Washington, DC, Rep. DOE/EIA-0383(2008), ,* Jun. 2008.

[2] U.S. Environ Protection Agency, "Inventory of U.S. greenhouse gas emissions and sinks," *1990–2006, Washington, DC, ,* Apr. 15, 2008. 2006.

[3] http://www.connectingcommuters.org/carpool/, Accessed Dec. 2015.

[4] M. Menendez and C. F. Daganzo, "Effects of HOV lanes on freeway bottlenecks," *Transportation Research Part B: Methodological,* vol. 41, pp. 809-822, 10// 2007.

[5] J. Kwon and P. Varaiya, "Effectiveness of California's High Occupancy Vehicle (HOV) system," *Transportation Research Part C: Emerging Technologies,* vol. 16, pp. 98-115, 2// 2008.

[6] A. Guin, M. Hunter, and R. Guensler, "Analysis of Reduction in Effective Capacities of High-Occupancy Vehicle Lanes Related to Traffic Behavior," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2065, pp. 47-53, 2008.

[7] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," presented at the Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, 2010.

[8] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, 2011.

[9] A. Aziz, K. Sanwal, V. Singhal, and R. Brayton, "Model-checking continuous-time Markov chains," *ACM Trans. Comput. Logic,* vol. 1, pp. 162-170, 2000.