Taxi data in New York city: A network perspective

Conference Paper · November 2015

DOI: 10.1109/ACSSC.2015.7421468

2 authors:

Joya Deri
Carnegie Mellon University
10 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Jose M F Moura
Carnegie Mellon University
584 PUBLICATIONS **10,047** CITATIONS

SEE PROFILE

TAXI DATA IN NEW YORK CITY: A NETWORK PERSPECTIVE

Joya A. Deri[1] *and José M. F. Moura*[1,2]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[2]Visiting, Center for Urban Science and Progress (CUSP), New York University, NYC, NY 11201
E-mail: {jderi,moura}@andrew.cmu.edu

ABSTRACT

We work with the "NYC Taxi Data Set," a historical repository of 750 million rides of taxi medallions over a period of four years (2010–2013). This data set provides rich (batch) information on the movements in an urban network as its citizens go about their daily life. We present a spectral analysis of taxi movement based on the graph Fourier transform, which necessitates the spectral decomposition of a large directed, sparse matrix. Important considerations toward handling this matrix are discussed. Preliminary results show that our method allows us to pinpoint locations of co-behavior for traffic in the Manhattan road network.

1. INTRODUCTION

Recent open-data initiatives such as the New York City Open Data Project provide access to large ("big") data sets that allow for the analysis of non-traditional "signals" on an urban framework. In this paper, we consider the NYC Taxi Data Set, which consists of details for around 750 million rides of taxi medallions from 2010 to 2013. Each record contains about thirty fields of data including number of passengers, start- and end-time stamps, start- and end-GPS location coordinates, amount paid itemized as fare, tax, and tip, as well as (coded) id information on the driver, owner, and vehicle. This data set provides static details for each taxi ride; however, we seek a dynamic representation that incorporates the movement of the taxis as constrained by an underlying road network. In particular, we seek to represent the given data as a graph signal in order to identify spatial trends.

A key aspect of a road network, particularly in New York City, is the presence of one-way as well as two-way streets. Such physical, man-made constraints can be captured by the asymmetric adjacency matrix A of a directed graph $\mathcal{G}(V, A)$, where V is the set of $N = |V|$ nodes. We define the *graph Fourier transform* $\widehat{s} \in \mathbb{C}^N$ of a signal $s \in \mathbb{C}^N$ over $\mathcal{G}$ as in [1]:

$$\widehat{s} = V^{-1}s, \tag{1}$$

where $V \in \mathbb{C}^{N\times N}$ is the (generalized) eigenvector matrix in the Jordan decomposition of the adjacency matrix

$$A = VJV^{-1}. \tag{2}$$

In cases where the undirected graph is an appropriate model, frameworks that use the symmetric adjacency matrix [2] or the graph Laplacian [3, 4, 5] are also applicable. Here we use the formulation in (5) to handle the directed road network.

Applying the graph signal processing framework for a directed graph based on a real-world system creates numerous challenges. In particular, for non-diagonalizable, or *defective*, matrices, the presence of eigenvalues of high algebraic multiplicity complicates the computation of the Jordan decomposition, which we discuss in detail in Section 4. Previous work on identifying eigenvalues of high multiplicity and Jordan forms for defective matrices can be found in [6, 7, 8, 9, 10].

The rest of the paper is as follows. We describe the data set and graph signals we consider in Section 2. Section 3 presents the graph signal processing framework while Section 4 describes necessary steps to account for the defectiveness of the underlying road network. Section 5 presents preliminary results, and we conclude in Section 6.

2. NETWORK AND GRAPH SIGNAL

We first present the road network we use in Sections 4 and 5. We also describe how we construct graph signals from the New York City taxi data.

2.1. Road Network

We define a 6408×6408 directed road network $\mathcal{G} = \mathcal{G}(V, A)$ representing the Manhattan roads with node set V and the adjacency matrix $A \in \mathbb{R}^{6408\times 6408}$. A node $v \in V$ represents a spatial location either at an intersection or at a point along a road, where the points were determined using the data portal in [11]. The adjacency matrix A is an asymmetric binary (0/1) matrix where a nonzero element at A_{ij} indicates the existence of a directed edge from node v_i to node v_j. Such a directed edge exists if traffic is legally allowed to move from v_i to v_j and can be determined from Google Maps or OpenStreetMap. The network is strongly connected with 14,418 directed edges.

In order to define certain graph signals, we use a strongly connected road network $\mathcal{G}_{\text{all}} = \mathcal{G}(V_{\text{all}}, A_{\text{all}})$ to represent the

This work has been supported by NSF grants CCF1011903 and CCF1513936.

roads of all New York City boroughs and airport areas, where the node and edge sets contain those of the Manhattan network $\mathcal{G}$; that is, $V \subset V_{\text{all}}$ and $E \subset E_{\text{all}}$, where E and E_{all} are the edge sets for $\mathcal{G}$ and $\mathcal{G}_{\text{all}}$, respectively. The network $\mathcal{G}_{\text{all}}$ has 79,234 nodes and 224,966 directed edges.

2.2. Taxi Data

We use the New York City taxi data [12] to define signals over the Manhattan grid $\mathcal{G}$. Each entry in the data set corresponds to one of about 750 million taxi rides. The key data fields we consider for each ride are the start- and end-geocoordinates (latitude and longitude), and the start- and end-time stamps. We ignore rides that have infeasible geocoordinates, e.g., if the geocoordinates correspond to a point in the middle of a body of water or a city park. We also impose that the ride distance must be greater that 0.2 miles and that the ride duration must be greater than one minute. More considerations for determining valid rides can be found in [13]. We have about 700 million valid taxi rides after cleaning the data.

We map a geocoordinate x (either a start- or an end-point of a ride) to a node $v \in V$ on the road network in the following way. First, the set of nodes $W \subset V$ that lie within a small radius of x is found. For each $w \in W$ and each edge $e \in E$ such that w is an endpoint of e, the orthogonal projection x_e from x to the line that represents edge e is found, and the Euclidean distance $d(x_e, x)$ from x_e to x is calculated. The mapped point $v \in V$ corresponds to the endpoint of the edge e^* that is closest to x, where e^* minimizes $d(x_e, x)$.

2.3. Graph Signal

We would like to construct graph signals $s \in \mathbb{R}^{6408}$ that capture the movement of the taxis on the Manhattan grid. We do so by computing Dijkstra shortest paths on $\mathcal{G}(V, A)$ for each ride; that is, if a ride i starts at mapped coordinate $x_i \in V$ and ends at $y_i \in V$, we compute the Dijkstra shortest path $(x_1, x_2, \ldots x_p)$ where p is the path length. Then each node of the path would contribute to the corresponding p elements of s_i.

In particular, we define each element of s as the average number of taxi rides that pass through node $v_i \in V$ per hour. Let t be the time index that represents an hour in a week. Then we let P_t be the set of paths that start or end at time t, $P_{t,j}$ be the jth path in P_t with path length p_j, and $P_{t,j,k}$ be the kth node of the jth path in P_t. We define the elements $s_{t,i}$ of the graph signal $s_t \in \mathbb{R}^{6408}$ as

$$s_{t,i} = \frac{1}{|P_t|} \sum_{j=1}^{|P_t|} \sum_{k=1}^{p_j} \chi_{v_i} \left(P_{t,j,k} \right), \tag{3}$$

where $\chi_v(w)$ is the indicator function that is 1 if $w = v$ and 0 otherwise. Equation (3) computes the average number of times a taxi ride passes through a node $v_i \in V$ at time t.

In Section 5, we will use graph signals defined as (3) to represent the average number of trips per unit time over the Manhattan grid. These signals will be analyzed via the graph Fourier transform as we discuss in Section 3.

3. GRAPH SIGNAL PROCESSING FRAMEWORK

In order to perform a frequency analysis of a graph signal, we need a definition of frequency and frequency order, and we need a transform that allows the signal to be represented as a sum of frequency components. We use the framework presented in [1, 14, 15] to formulate both parts. We consider the eigenvalues of the adjacency matrix A to be the frequencies, or modes. The eigenvalues may be complex for asymmetric A, so we define frequency order in terms of the total variation of the corresponding eigenvectors as discussed in [15].

We define the Fourier transform matrix $F \in \mathbb{C}^{N \times N}$ on the network $\mathcal{G}(V, A)$ to be the inverse of the eigenvector matrix $V \in \mathbb{C}^{N \times N}$ obtained by the Jordan decomposition of A given by $A = VJV^{-1}$ [1, 14, 15], or

$$F = V^{-1}. \tag{4}$$

It is important that the eigenvectors in V be sorted by the frequency order. For a graph signal $s \in \mathbb{R}^N$, its *Fourier transform* with respect to the graph $\mathcal{G}$ is then

$$\widehat{s} = V^{-1} s, \tag{5}$$

and the *inverse Fourier transform* is

$$s = V\widehat{s}. \tag{6}$$

We note that V has full rank (its inverse exists), but it may not be orthogonal if A has eigenvalues with algebraic multiplicity greater than one. In Section 4, we identify problems that arise with the presence of multiple eigenvalues and discuss potential solutions.

4. SPECTRAL DECOMPOSITION FOR SPARSE, DIRECTED NETWORKS

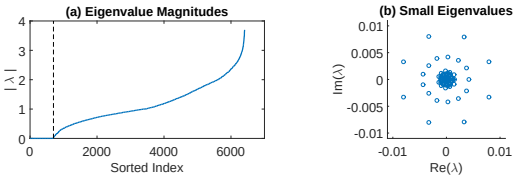

Fig. 1. (a) Eigenvalue magnitudes of the road network adjacency matrix in ascending order. The magnitudes are "close" to zero in the index range 1 to 699 (up to the dotted line). (b) Eigenvalues in the sorted index range 1 to 699 plotted on the complex plane.

We will describe details for computing the eigenvector matrix V in the Jordan decomposition of A for the case of a non-diagonalizable adjacency matrix $A \in \mathbb{R}^{N \times N}$. Our motivation is the directed road network $\mathcal{G} = \mathcal{G}(V, A)$ from Section 2.1. For this particular adjacency matrix, we first observe that the eigenvector matrix $V_{\text{obs}} \in \mathbb{C}^{N \times N}$ generated by a standard eigenvalue solver as in MATLAB is not full

rank, where the rank can be computed with either the singular value decomposition or the QR decomposition; i.e., A is not diagonalizable. We thus require the Jordan form to apply Equation (5). For details about the Jordan form, we direct the reader to [16].

We find that the dimension of the null space of A (the geometric multiplicity of the zero eigenvalue) is 446. Furthermore, we notice that there are 699 eigenvalues of magnitude close to zero as seen in Figure 1(a). If these eigenvalues represent true zero eigenvalues, the algebraic multiplicity of the zero eigenvalue would be 699, and we need to find $699 - 446$ or 253 generalized eigenvectors to complete the Fourier basis.

We remark that there is ambiguity in determining whether the 699 eigenvalues of small magnitude represent true zero eigenvalues. Accurately identifying eigenvalues of high algebraic multiplicity can be a hard problem – see, for example, [6, 7, 8, 9, 10]. Furthermore, we observe that these eigenvalues of small magnitude form constellations with centers close to zero, as shown in Figure 1(b). As discussed in [6, 7], the presence of such constellations may indicate the presence of an eigenvalue of high algebraic multiplicity; reference [6] states that an approximation of this eigenvalue may be the center of the cluster of numerical eigenvalues, while [7] presents an iterative algorithm to approximate a "pseudoeigenvalue." It is unclear[1] whether the constellations we observe are a result of the multiplicity of a zero eigenvalue or whether they are the actual eigenvalues of A. In addition, we do not know if the presence of true eigenvalues are hidden within the constellation, i.e., whether an eigenvalue $\lambda \neq 0$ exists in this set of small eigenvalues. In the face of this uncertainty, we outline the steps we took to compute a full-rank eigenvector matrix V that we can use to define the graph Fourier transform in Equation 5. Section 4.1 discusses the construction of the Jordan chains for the zero eigenvalue, and Section 4.2 explains the case for considering non-zero but small eigenvalues to extend the Fourier basis.

4.1. Computing Jordan chains for the zero eigenvalue.

We first compute Jordan chains for the zero eigenvalue. Since the maximum Jordan chain length is unknown, we need an approximation. For this, we use Definition 1 in [9]:

Definition 1 ([9]). *Consider a matrix $A \in \mathbb{C}^{N\times N}$ with singular values $\sigma_1(A) > \cdots > \sigma_N(A)$ that is scaled so that $\sigma_1(A) = 1$. Let $m_k = \left|\mathcal{N}(A^k)\right|$ denote the dimension of the null space of A^k . In addition, let α and δ be positive constants; δ is usually on the order of machine precision and α is significantly greater than δ. Then 0 is a numerically multiple eigenvalue with respect to α and δ if*

$$\sigma_{N-m_k}\left(A^k\right) > \alpha > \delta > \sigma_{N-m_k+1}\left(A^k\right), \qquad (7)$$

for $k = 1, 2, \ldots, h$, where h is the maximum Jordan chain length for the zero eigenvalue.

Since the constants α and δ have different orders of magnitude, Equation (7) implies that singular value $\sigma_{N-m_k}(A^k)$ is significantly greater than $\sigma_{N-m_k+1}(A^k)$.

Definition 1 serves two purposes for our application. It first verifies the existence of a numerical zero eigenvalue. It also tells us that a value of k at which Equation (7) fails cannot be the maximum Jordan chain length of the zero eigenvalue. This implies the following method to find the maximum Jordan chain length h: increment the value of k starting from $k = 1$, and let $k = k'$ be the first value of k such that Equation (7) fails. Then the maximum Jordan chain length for eigenvalue zero is $h = k' - 1$.

We apply Definition 1 to the adjacency matrix A of the Manhattan road network and obtain Table 1. The columns correspond to the power k of A, the dimension m_k of the null space of A^k, the index $N - m_k$ of the first singular value of A^k we examine, and the values of the singular values at indices $N - m_k$ and $N - m_k + 1$. The machine precision on the computers we use is on the order of 10^{-16}, and we see that we can find a δ on the order of 10^{-15} or 10^{-16} as well as a significantly larger constant α such that the inequality (7) holds for $k \in [1, 4]$. The inequality begins to fail at $k = 5$; thus, we expect the maximum numerical Jordan chain length to be no more than 3 or 4.

k	m_k	$N - m_k$	$\sigma_{N-m_k}\left(A^k\right)$	$\sigma_{N-m_k+1}\left(A^k\right)$
1	446	5962	1.9270×10^{-3}	1.2336×10^{-15}
2	596	5812	2.1765×10^{-6}	6.9633×10^{-16}
3	654	5754	1.4013×10^{-8}	3.4250×10^{-16}
4	678	5730	1.1853×10^{-10}	3.1801×10^{-16}
5	692	5716	2.0163×10^{-11}	8.4063×10^{-14}
6	700	5708	9.6533×10^{-11}	8.2681×10^{-11}

Table 1: Singular values to validate existence of a numerical zero.

We find the eigenvectors for eigenvalue zero as follows. First we compute the null space $\mathcal{N}_A$ of A to find the corresponding eigenvectors. Each of these eigenvectors corresponds to a Jordan block in the Jordan decomposition of A, and the Jordan chains with maximum length h (as determined by Definition 1) are computed by the recurrence equation [16]

$$Av_k = \lambda v_k + v_{k-1} = v_{k-1}, \qquad (8)$$

where $k \in [1, h]$ and $v_0 \in \mathcal{N}(A)$. If the number of linearly independent proper and generalized eigenvectors equals N, we are done; otherwise, we need to extend the Fourier basis as described in the next section.

We remark that it is important to generate the Jordan chains from the vectors in the null space of $\mathcal{N}(A^h)$, where h is the maximum Jordan chain length. Then the Jordan chain can be constructed by direct application of the recurrence equation (8). Finding the Jordan chains starting from the

[1] Considering double-precision floating point (64 bit) and that the number of operations to compute the eigenvalues of an $N \times N$ matrix is $O(N^3)$, we expect our precision to be on the order of 10^{-6} or 10^{-7}. We observe numerous eigenvalues in Figures 1(a) and (b) that have this order of magnitude.

proper eigenvectors is numerically unstable and may lead to solving inconsistent systems of equations [17].

4.2. Extending the Fourier basis.

If the eigenvectors, including those found for the zero eigenvalue, do not span $\mathbb{C}^N$, we need to determine nonzero eigenvalues and their eigenvectors to complete the Fourier basis. To do so, we sort the small eigenvalues by decreasing magnitude to partition them into clusters. Then, for each cluster, we try two options. We estimate the center λ_{ctr} of the cluster and compute the corresponding eigenvectors. We keep these eigenvectors if they extend the Fourier basis. Otherwise, we treat each eigenvalue in the cluster as a true eigenvalue, compute the eigenvectors, and keep the eigenvectors that extend the Fourier basis.

We use the methods discussed in Sections 4.1 and 4.2 to obtain a full rank matrix $V \in \mathbb{C}^{6408\times 6408}$ of proper and generalized eigenvectors. The maximum Jordan chain length we find is two, which is consistent with our findings in Table 1. The inverse of V is the graph Fourier transform matrix (see Section 3), which we use for our analysis in the next section.

5. EMPIRICAL RESULTS

We do a Fourier analysis of the average number of taxis trips in the Manhattan network on Sundays from 8am to 9am, averaged over four years for the months of June, July, and August. The magnitudes of the Fourier coefficients in Figure 2 show that the road network acts as a low-pass filter on the signal. Furthermore, we observe the highest frequency coefficient magnitudes at frequency indices 107, 108, 283, 4386, and 4387 corresponding to (proper) eigenvectors of eigenvalues $0.0033 + 0.0079i$, $0.0033 - 0.0079i$, 0.0366, $1.3793 + 0.0007i$, and $1.3793 - 0.0007i$, respectively.

To get a sense of the intuition behind the highly expressed graph Fourier coefficients, we plot the component magnitudes for the eigenvector that has the Fourier coefficient of maximum magnitude (frequency index 107). Fig. 3 shows six highly expressed (complex-valued) components, which we plot on a map of Manhattan, NYC, in Fig. 4. The topmost coordinates, corresponding to the largest eigenvector components in Fig. 3, are located on the on-ramp of Henry Hudson Parkway at 79th Street. The next largest components are the bottom-most coordinates of Fig. 4, located in the Financial District close to Hanover Square. The last highly expressed coordinates are in Little Italy.

From Figures 3 and 4, we observe that an eigenvector specifies locations in the city, whether adjacent or non-adjacent, that are related in a spectral sense. For an eigenvector corresponding to a low frequency, such as the one plotted in Figure 3, these locations are sparse. Furthermore, the high expression of an eigenvector as in Figure 2 indicates that the graph signal is highly influenced by the locations corresponding to the eigenvector's significant components.

We see that the spectral analysis provided by the graph Fourier transform allows for the identification of traffic

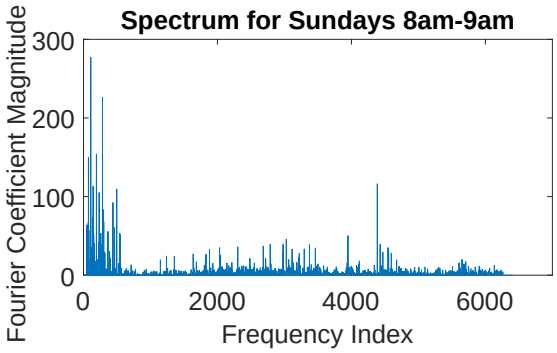

Fig. 2. Spectrum for the graph signal of average trips on Sundays from 8am to 9am. Small frequency indices correspond to low frequencies while high frequency indices correspond to high frequencies, as discussed in Section 3. The Fourier coefficient at frequency index 107 has the largest magnitude.

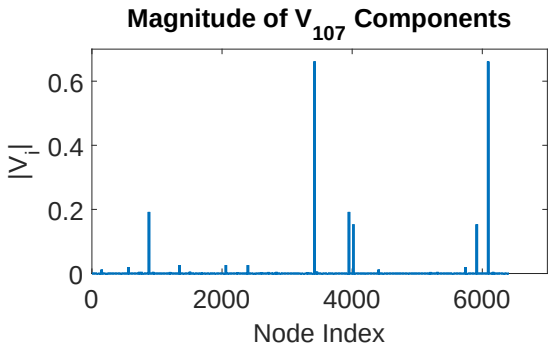

Fig. 3. Magnitude of the components of eigenvector 107. There are six highly expressed components.

Fig. 4. Map showing the highly expressed components of eigenvector 107 (plotted with [18]). The six locations of interest pair off into three clusters.

"hotspots" that highlight locations of significant co-behavior over the Manhattan grid. Such analyses may be useful for urban planning and anomaly detection.

6. CONCLUSION

This paper presents a spectral analysis of New York City taxi data via the graph Fourier transform. Besides the computational issues posed by the size of the data, we must also handle an underlying road network that is large, directed, and sparse. We discuss a way to perform the spectral (Jordan) decomposition for such a matrix with the example of the Manhattan road network. Our preliminary results show that the analysis allows us to pinpoint locations in New York City that exhibit co-behavior at a certain time of day and a certain day of week.

7. ACKNOWLEDGEMENTS

We would like to thank Professor Zhonggang Zeng from the Department of Mathematics at Northeastern Illinois University for pointing us to reference [7].

8. REFERENCES

[1] A. Sandryhaila and J.M.F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 80–90, Aug. 2014.

[2] B. Miller, N.T. Bliss, and P.J. Wolfe, "Toward signal processing theory for graphs and non-Euclidean data," in *Proceedings of the 2010 IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 5414–5417.

[3] D. Shuman, S.K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, Apr. 2013.

[4] X. Zhu and M. Rabbat, "Approximating signals supported on graphs," in *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2012, pp. 3921–3924.

[5] S.K. Narang and A. Ortega, "Perfect reconstruction two-channel wavelet filter banks for graph structured data," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2786–2799, Jun. 2012.

[6] Z. Zeng, "Regularization and Matrix Computation in Numerical Polynomial Algebra," in *Approximate Commutative Algebra*, L. Robbiano and J. Abbott, Eds., Texts and Monographs in Symbolic Computation, pp. 125–162. Springer Vienna, 2010.

[7] Z. Zeng, "Sensitivity and Computation of a Defective Eigenvalue," preprint at `http://orion.neiu.edu/~zzeng/Papers/eigit.pdf`, Apr. 2015.

[8] L.N. Trefethen and M. Embree, *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*, Princeton University Press, 2005.

[9] A. Ruhe, "An algorithm for numerical determination of the structure of a general matrix," *BIT Numerical Mathematics*, vol. 10, no. 2, pp. 196–216, 1970.

[10] G. H. Golub and J. H. Wilkinson, "Ill-Conditioned Eigensystems and the Computation of the Jordan Canonical Form," *SIAM Review*, vol. 18, no. 4, pp. 578–619, 1976.

[11] Baruch College: Baruch Geoportal, "NYC Geodatabase," URL: `https://www.baruch.cuny.edu/confluence/display/geoportal/NYC+Geodatabase`, Accessed 31-Aug.-2014.

[12] B. Donovan and D. Work, "New York City Taxi Data 2010-2013," URL: `https://uofi.box.com/s/zmggziub40wx1bq2h9bq/`, Accessed 31-Aug.-2014.

[13] B. Donovan and D. Work, "Using coarse GPS data to quantify city-scale transportation system resilience to extreme events," in *arXiv preprint arXiv:1507.06011*, Jul. 2015.

[14] A. Sandryhaila and J.M.F. Moura, "Discrete signal processing on graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[15] A. Sandryhaila and J.M.F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014.

[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, New York, NY, USA: Academic, 2nd ed. edition, 1985.

[17] J.H. Kwak and S. Hong, *Linear Algebra*, Birkhäuser Boston, 2004.

[18] "Mapcustomizer.com," URL: `http://www.mapcustomizer.com`, Accessed 07-Nov.-2015.